Project Document Cover Sheet

Project Information			
Project Acronym	IDMB		
Project Title	Institutional Data Management Blueprint		
Start Date	1 Oct 2009	**End Date**	31 Mar 2011
Lead Institution	University of Southampton		
Project Director	Dr Kenji Takeda		
Project Manager & contact details	Clint Styles, School of Engineering Sciences, University of Southampton C.Styles@soton.ac.uk, Tel. 023 8059 3601		
Partner Institutions	Digital Curation Centre, University of Oxford, National Oceanography Centre		
Project Web URL	http://www.southamptondata.org		
Programme Name (and number)	Managing Research Data (Research Data Management Infrastructure)		
Programme Manager	Simon Hodson		

Document Name			
Document Title	*Initial Findings Report*		
Reporting Period	1 Oct 2009 to 31 Aug 2010		
Author(s) & project role	Dr Kenji Takeda, Project Director		
Date	15 Sep 2010	**Filename**	IDMBInitialFindingsReportv4.doc
URL	*if document is posted on project web site*		
Access	☒ Project and JISC internal	☐ General dissemination	

Document History		
Version	**Date**	**Comments**
1	15 Sep 2010	Draft for review by project team
2	22 Sept 2010	Draft for review by Steering Group
3	15 Nov 2010	Final version for review by Steering Group and team
4	15 Dec 2010	Final version for distribution

University of Southampton

Institutional Data Management Blueprint Project

Initial Findings Report

September 2010

Project Director: Dr Kenji Takeda

Executive Summary

In the 10th anniversary year of the Open Archiving Initiative it is necessary to elevate research data to be a first-class citizen in the world of open scholarly communication. Such a profound goal requires far more than technical capability, but encompasses significant change for all stakeholders. Data curation and data management is often seen as an additional task for researchers. It is, however, a critical part of research best practice. In this project we are attempting to make it a seamless part of a researchers' daily workflow across a wide range of disciplines as a cornerstone of research practice.

This report describes the initial findings from the Institutional Data Management Blueprint (IDMB) project, which aims to create a practical and attainable institutional framework for managing research data throughout its lifecycle that facilitates ambitious national and international e-research practice. The objective is to produce a framework for managing research data across the whole lifecycle that encompasses a whole institution (exemplified by the University of Southampton) and based on an analysis of current data management requirements for a representative group of disciplines with a range of different data.

This report covers the data management audit, kick-off workshop, and data management framework development within the project.

The project website is at www.southamptondata.org

Key Findings

Notable conclusions so far include:

- There is a need from researchers to share data, both locally and globally;
- Data management is carried out on an ad-hoc basis in many cases;
- Researchers' demand for storage is significant, and outstripping supply;
- Researchers resort to their own best efforts in many cases, to overcome lack of central support;
- Backup practices are not consistent, with users wanting better support for this;
- Researchers want to keep their data for a long time;
- Data curation and preservation is poorly supported;
- Schools research practice is embedded and unified;
- Schools data management capabilities vary widely.

In terms of gap analysis, the following major conclusions can be inferred:

- Policy and governance is robust, but is not communicated to researchers in the most accessible way;
- Services and infrastructure are in place, but lack capacity and coherence;
- There is a lack of training and guidance on data management.

It is apparent that there is no coherent data management approach, with the current business model not being scalable, nor sustainable, to meet the current and future demand required to support the university's strategic goals to deliver research excellence.

A three-layer metadata strategy based on Dublin Core has been proposed to provide a unified approach to improving data management across all disciplines.

Three pilot implementations around archaeology, the Southampton Nanofabrication Centre, and meta-search across federated repositories, have been described and development work is starting on these.

It is clear that the current data management situation at the University of Southampton is analogous to the HPC landscape at Southampton a decade ago. The institution successfully moved to a more coordinated HPC framework since then that provides world-leading capability to researchers through a sustainable business model. A similar step change in data management capability is required in order to support researchers to achieve the University's ambitious strategic aims in the coming decade.

Recommendations

The data management audit and gap analysis indicates where improvements can be made in the short, medium and long-term to improve data management practices and capabilities at the University. The following preliminary recommendations are put forward for short (one year), medium (one to three years), long (more than three years) term action. The exact timing of implementation of recommendations is subject to further prioritization by the institution.

Short-term (one year)

Crucial to supporting researchers is the consolidation of data management into a coherent framework that is easy to understand, use, and has a sustainable business model behind it. A number of major recommendations are put forward here for the short-term:

- Create an institutional data repository
- Develop a scalable business model
- One-stop shop for data management advice and guidance

Medium-term (one to three years)

The medium term (1-3 years) presents opportunities to enhance research capability and profile:

- Comprehensive and affordable backup service for all
- Open research data mandate, and supporting infrastructure
- Research data lifecycle management
- Embedding data management training and support

Long-term (more than three years)

Long-term aspirations can provide significant benefits realisation across the whole University, and a stable foundation for the future:

- Provide coherent data management support across all disciplines
- Embed exemplary data management practice across the institution
- Agile business plan for continual improvement

Acknowledgements

The Institutional Data Management Blueprint project is funded by the UK's Joint Information Systems Committee (JISC). It is part of the Managing Research Data programme, managed by Simon Hodson.

The project investigators are:

- Kenji Takeda (Engineering Sciences, Project Lead)
- Mark Brown (University Librarian)
- Simon Coles (Chemistry)
- Les Carr (ECS)
- Graeme Earl (Archaeology)
- Jeremy Frey (Chemistry)
- Peter Hancock (iSolutions)

With the project team including:

Project Manager

- Clint Styles

Library team

- Wendy White
- Fiona Nichols
- Michael Whitton
- Harry Gibbs
- Christine Fowler
- Pam Wake

iSolutions

- Steve Patterson

The authors would like to thank the project steering group for their input and time:

- Adam Wheeler (Provost and Deputy Vice-Chancellor)
- Philip Nelson (Deputy Vice-Chancellor, Research)
- Graham Pryor (Digital Curation Centre)
- Sally Rumsey (University of Oxford)
- Helen Snaith (National Oceanography Centre Southampton)
- Simon Cox (Engineering Sciences, EPSRC HPC Technology Watch Panel)
- Peter Hancock (iSolutions, Director)
- Kenji Takeda (Engineering Sciences)
- Mark Brown (University Librarian)
- Jeremy Frey (Chemistry)

Table of Contents

Nomenclature

AIDA	Assessing Institutional Data Assets
ECS	School of Electronics and Computer Science
HPC	High Performance Computing
IDMB	Institutional Data Management Blueprint Project
iSolutions	Central IT organisation
JISC	Joint Information Systems Committee
NAS	Network attached storage
SES	School of Engineering Sciences
SESNET	School of Engineering Sciences network
SUSSED	University web portal
UoS	University of Southampton
VPN	Virtual Private Network

1 Introduction

In the run-up to the 10th anniversary of the Open Archiving Initiative it is necessary to elevate research data to be a first-class citizen in the world of open scholarly communication. Such a profound goal requires far more than technical capability, but encompasses significant change for all stakeholders. The aim of the Institutional Data Management Blueprint (IDMB) project is to create a practical and attainable institutional framework for managing research data that facilitates ambitious national and international e-research practice. The objective is to produce a framework for managing research data that encompasses a whole institution (exemplified by the University of Southampton) and based on an analysis of current data management requirements for a representative group of disciplines with a range of different data. Building on the developed policy and service-oriented computing framework, the project will scope and evaluate a pilot implementation plan for an institution-wide data model, which can be integrated into existing research workflows and extend the potential of existing data storage systems, including those linked to discipline and national shared service initiatives.

The project builds upon a decade of previous open access repository initiatives at Southampton to create a coherent set of next actions for an institutional, cross-discipline 10-year roadmap, which will be flexible in accommodating future moves to shared services, and provide a seamless transition of data management to knowledge transfer, from the individual to the community and from the desktop to institutional, national and international repositories (Figure 1). The outcomes from this project, which will draw together technical, organisational and professional expertise from across the institution, will be widely disseminated within the sector as a form of HEI Data Management "Business Plan How-To". Through this project the University of Southampton will continue its work at the forefront of the Open Access movement and leverage its innovation and experience in this area to the benefit of the UK and global academic community.

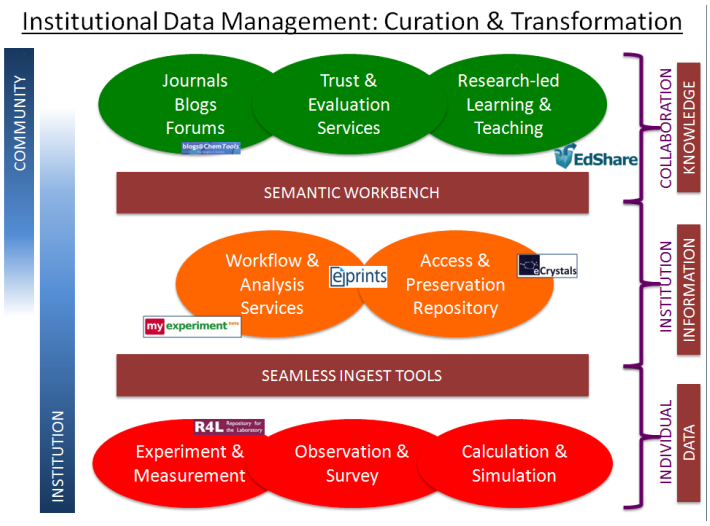

Figure 1. Institutional Data Management and Transformation

Since the Budapest Declaration in 2001[1], the Open Access movement has seen scientific knowledge become more available, with the community starting to experience the benefits of this, in terms of better awareness and enhanced citations. In order to fully realise the benefits of open access, the next step is to make the data upon which this knowledge is based, and the processes and analyses by

[1] http://www.soros.org/openaccess

which it is attained, more freely available and easy to access. The successful management, curation and preservation of UK research data has been increasingly recognised as a significant issue for the national research infrastructure since the appearance of the government report *"Science and Innovation Investment Framework 2004-2014"* [2], which in turn has led to a series of studies and reports designed to help define the research data landscape.[3]. There has been a great deal of work contributed to defining and scoping aspects of the research data lifecycle, a number of which have sought to engage directly with researchers, which is recognised as increasingly important[4]. Defining the responsibilities for managing data from inception to preservation is now clearly recognised as a complex process shared between individual researchers and research groups, institutions, funders and national agencies. This is driven by many agendas, including groups of users, different funding agencies and programmes, politics, and technology trendsetters. A constant factor is the institution - a centre for cohesion, curation and cooperation - which is responsible for its own research data at some, or maybe all, of its lifetime, within a fragmented and volatile world. In order to acknowledge and manage these responsibilities, institutions require an overall framework within which to plan and develop their data management strategy. Many of the landscape studies so far have been highly detailed analytical descriptors of theoretical models, with some testing of assumptions, which institutions can find difficult to implement, and which can be too complex to win engagement from researchers. The management of data requires a multifunctional team approach which can bring together the knowledge and expertise of both researchers and professionals within an institutional policy and technical framework. Southampton has a proven track record in creating a team approach to managing research outputs evidenced from the extensive work with repositories in institutions and disciplines over the past decade, as shown in Figure 2 (taken from http://bit.ly/25Cght).

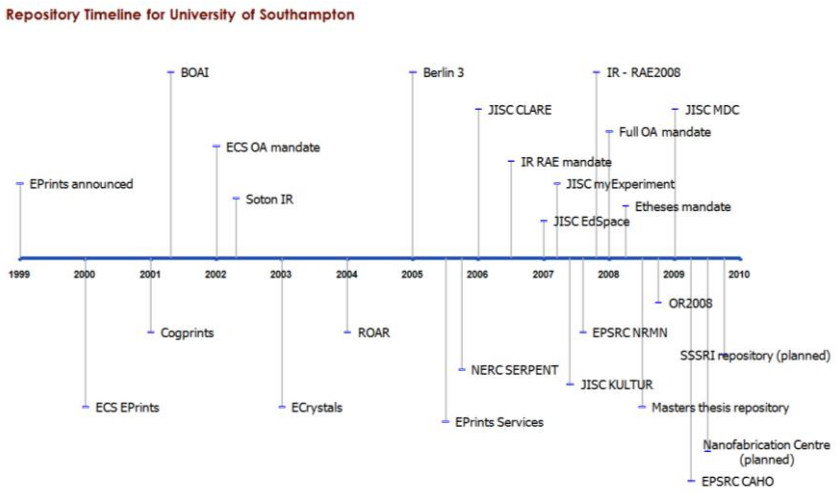

Figure 2: Repository Initiatives at Southampton. http://bit.ly/25Cght

This project focuses on developing a long-term solution, fully detailed and costed, for a single institution by leveraging open standards and service-oriented approaches. The work will involve developing policy, distilling best practice from an institutional viewpoint, deploying a pilot data management framework, and dissemination within and outside the University of Southampton. One of the principal aims of the project is to provide a framework which is deliverable in terms both of cost and practical application, and which is the result of working directly with researchers to avoid over-complex and potentially time consuming processes. An enterprise architecture will be developed based on a service-oriented approach that encompasses existing services, and is extensible as new

[2] HMSO 2004
[3] most notably the feasibility study for a distributed data service, the UKRDS, see http://www.ukrds.ac.uk/
[4] http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/dataauditframework.aspx

services become available; in line with the JISC e-Framework. This is based on the extensive experience of the Southampton project team, through a number of data management and related initiatives, funded from a variety of sources.[5] Recommendations from the DISC-UK DataShare project[6] highlights this level of detail as the next step – *"We propose that this collaborative effort across and among institutions can be a model for future development as we move from discussion to full implementation of policies and practices related to bringing data into repository environments"*. By going through the detailed planning process for one HEI we will provide a level of insight and experience that is currently not widely available within the sector for others to benefit from. We will use an 'open source' approach to strategy development, so that others can maximally benefit from this project.

Expected outcomes from the project include:

- Pathfinder for an institutional data management strategy for the next decade;
- Data management institutional blueprint based on an analysis of data management requirements and current best practice;
- Service-oriented, extensible enterprise architecture model for data management;
- 10-year business model roadmap;
- Best practice gap analysis report;
- Pilot implementation for infrastructure, human and technological;
- Workshops, training, website and reports for dissemination of best practice.

In achieving these outcomes, the project aims to add value to individuals, the institution and the research community, in the following ways:

- Coherent data management strategy for a single institution;
- Change management strategy for open access of data;
- Development of cross-professional skills base for managing research data, including graduate student training;
- Preservation and curation of research data at an institutional level;
- Advocacy and best practice guidance for research workflows and data across disciplines;
- Cost-benefit analysis of implementing the framework;
- Detailed business model blueprint for others, to accelerate early adoption.

For this project we take a holistic view, so our definition of data is very broad. It includes not only experimental, observational and derived research data, but also that required to produce, demonstrate and record provenance. Hence we include computer code and laboratory/field notes in our remit. This is to ensure that our institutional approach encompasses all of a researcher's requirements, which can then be narrowed as required.

[5] e.g. eCrystals Federation, DataShare, myExperiment, NERC SERPENT, Kultur, and Materials Data Centre amongst others.
[6] Green, A., Macdonald, S., Rice, R. (2009) "Policy-making for Research Data in Repositories: A Guide", DISC-UK DataShare

1.1 Report Structure

This report covers the initial findings from the Institutional Data Management Blueprint (IDMB) project. The following sections are included:

2. **Data management audit.** An audit has been carried out to find out how users manage their data, and how the University supports them. This is the results of an online questionnaire and interviews with 50 researchers. The AIDA (Assessing Institutional Digital Assets) toolkit has been used to benchmark current capability at the departmental/school and institutional level.
3. **Kick-off workshop**. A report on the kick-off workshop held on 24 March 2010, with over 40 attendees, describes the quick wins, long term dreams, and current issues described by the participants.
4. **Data Management Framework.** This section describes the current and proposed future direction, for an integrated data management framework for the University of Southampton. It includes: policy, governance and legal issues; services and infrastructure; gap analysis, and; metadata strategy.
5. **Pilot implementations.** A brief description of the three pilot studies that are being carried out in archaeology, the Nanofabrication Centre, and for a repository meta-search prototype.

The report concludes with a set of recommendations for the institution based on our initial findings.

Appendices are included with the questionnaire, interview questions, AIDA survey, and funders' policies.

2 Data Management Audit

2.1 Methodology

A key part of the IDMB project is to engage with users to ascertain current data management practice, support and constraints. In order to do this we have extended the data management audit carried out for the Southampton School of Social Sciences as part of the DataShare project[7]. This was carried out alongside face-to-face workshops with the research community, as described in Section 3.

The audit involved a four-pronged approach to gather quantitative and qualitative data at the individual, School and University levels:

- **Online questionnaire.** This was used to provide quantitative information across a spectrum of areas including current practice, policy, and governance. These were targeted at individuals in the Schools of Electronics & Computer Science, Engineering Sciences and Humanities.
- **Interviews.** Follow-up interviews with willing questionnaire respondents were used to obtain more details from individuals to provide more qualitative data. This allowed us to drill-down into specific area that participants were particularly interested/concerned with.
- **AIDA (Assessing Institutional Data Assets) tool[8].** The AIDA self-assessment tool is designed to provide benchmarking data on the level of data management capability available. Here it has been applied at the School and University levels.
- **Crowdsourcing.** In order to obtain additional feedback and experimental crowd-sourcing approach has been piloted. This is using the project website (www.southamptondata.org) and uses an *ideas box* approach.

In this section we describe the results from the questionnaire, interviews and AIDA benchmarking. Key findings are brought together in section 2.3. These findings are used to guide the data management framework described in section 4, the pilot implementations described in section 5, and produce initial recommendations in section 6.1.

[7] Gibbs, H., (2009), Southampton Data Survey: Our Experience and Lessons Learned, DISC-UK DataShare project

[8] http://aida.jiscinvolve.org/wp/

2.2 Results

2.2.1 Questionnaires

In order to obtain quantitative data around research data management, an online questionnaire was devised based on the approach used in the previous DataShare project. The project team refined the questionnaire to provide a balance of level of detail, completion time, and categorisation of results in the most usable format. The questionnaire comprises 30 questions and takes 15 or more minutes to complete.

The questionnaire was split into three sections: About You; About your data; Final comments. The full questionnaire is shown in Appendix I. While most of the questions were quantitative, additional text boxes for qualitative information were included where appropriate[9].

The survey and interviews required ethics clearance before they could be launched, and they have been successfully completed by participants in the Schools of Humanities, Electronics & Computer Science (ECS) and Engineering Sciences (SES). Following the kick-off workshop (see section 3), senior stakeholders agreed that the questionnaire should be rolled out across the whole University – comprising 23 Academic Schools. The final project report will include updated quantitative data and analysis for this more comprehensive survey.

Due to the ethics clearance, individual invitations were sent out to potential respondents to the questionnaire website (http://www.isurvey.soton.ac.uk/663) and the URL was advertised on the project website.

In total 114 researchers completed the questionnaire, of a total of 282 attempting it. This indicates a 40% completion rate, with the majority who did so achieving this in less than 15 minutes. The questionnaire may be streamlined for the university-wide rollout in an attempt to increase the completion rate. The breakdown of respondents by Academic School and research role is shown in Table 1.

Note that the answers to the questionnaire were from individuals, and the statements they make based on their own knowledge. Therefore some answers may not reflect the true situation, only the respondent's view.

[9] Where there is a mix of qualitative/quantitative data in a single question (e.g. 'other please specify options) this is marked with an *.

	Electronics and Computer Science	Engineering Sciences	Humanities (Archaeology only)	Total
Academic or equivalent (e.g. HEFCE-funded)	21	9	11	**41**
Research Fellow (e.g. Project-funded)	6	10	3	**19**
Research Student	19	10	16	**45**
Other	1	2	4	**7**
Total	**47**	**31**	**34**	**114**

Table 1. Survey respondents, broken down by school and role (Q1.2 and Q1.3)

2.2.1.1 About You

This section describes the responses for each question, including additional qualitative data where appropriate.

Q1.3 Research Role*

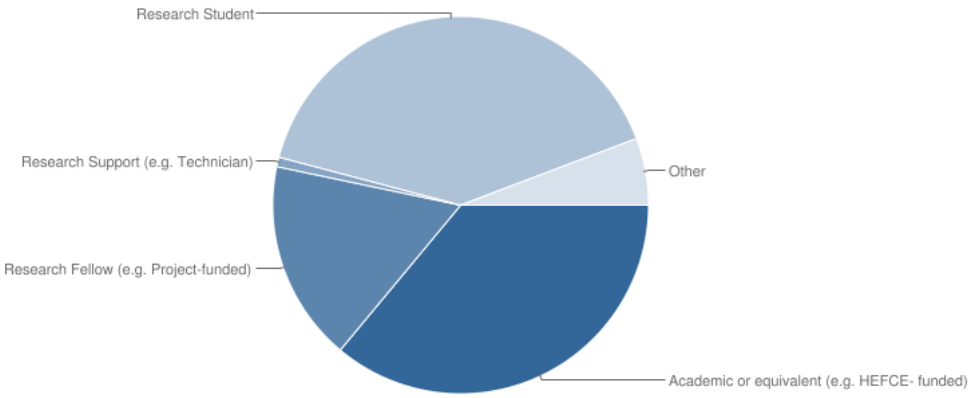

Figure 3. Please describe your research role.

A handful of respondents identified other research roles:

- Publicity coordinator (ECS)
- Research Engineer on KTP [Knowledge Transfer Partnership] Project (SES)
- Administration; Experimental Officer (Archaeology)

Q1.4 Area of Research*

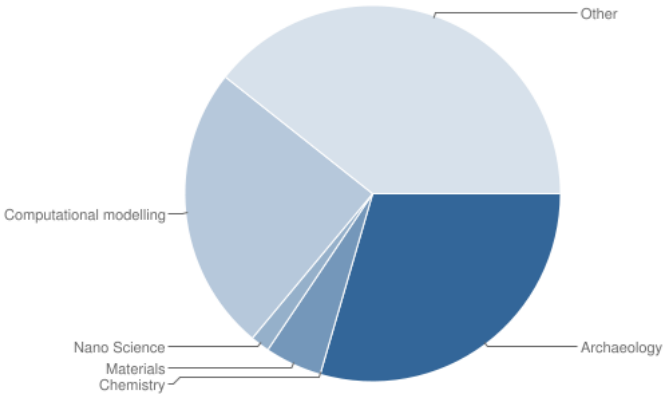

Figure 4. What is your area of research?

A wider range of researchers in ECS and SES were surveyed than originally considered when drafting the survey for ethical approval. Thus a large number identified their area of research as 'other'. Respondents described their research, which is diverse and difficult to categorise, as indicated in Table 2.

Electronics and Computer Science	Engineering Sciences
Communications (x 4) Web science / Semantic Web (x2) Power Engineering Computer Vision (x2) Medical Biological (x4) Human-Computer Interaction / Accessibility (x3) Networks / Grid (x2) Hydrodynamics / Fluid mechanics Image and sensor archives Industrial Information management e-learning (x2) Nano science, experimental engineering, Preservation and access to digital audio-visual content Software Engineering	Aerodynamics computational modelling experimentation in marine field Deployable structures Electrochemical Engineering Experimental Fluid Mechanics MEMS Ultrasonics Microfluidics Space Structures

Table 2. Questionnaire respondent research areas

2.2.1.2 About your data

Q2.1 Ownership of research data*

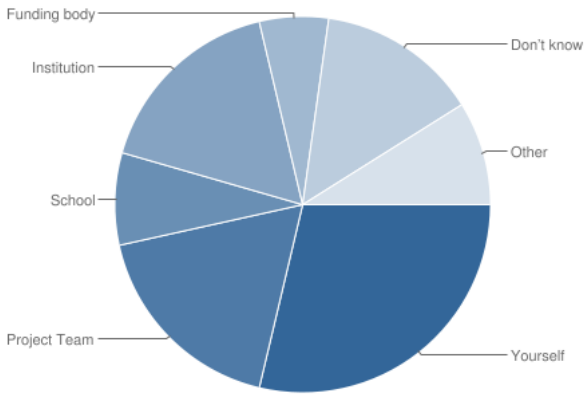

Figure 5. Who do you believe owns your research data?

The responses for this question, which could be multiple, show that 29% identify themselves as owning their data. This compares with 25% for School/University (combined) and only 6% for the funding body. 14% of respondents did not know who owned their research data.

A number of specific funding bodies were identified as owning data: PASCAL Network, Airbus France, Rolls Royce, Defence Science and Technology Laboratory and the BBC.

There were a number of comments indicating the ownership of their data varied depending on project agreements and could be shared between multiple bodies. A comment about archaeological objects indicated that they are the property of the owner until an agreement is reached for deposit in a museum

Q2.3 Data types*

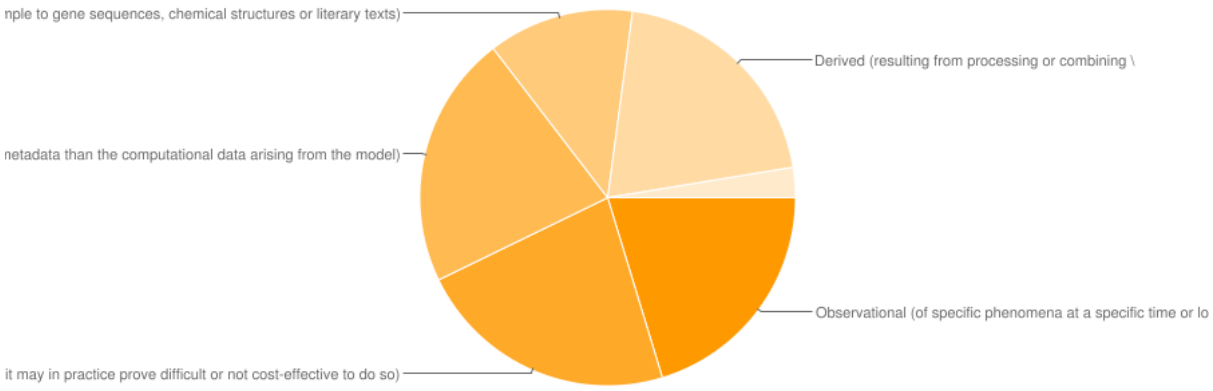

Figure 6. Characteristics of the data – please tick all that apply

The responses to this question are split relatively evenly between observational (20%), experimental (23%), computer code (22%) and derived data (20%). This is perhaps expected as there is typically a data pipeline going from experimental/observed/simulation data to derived data.

The split by School indicates that ECS and SES respondents held relatively more computer code (75% and 86% respectively) than in the Humanities (13%).

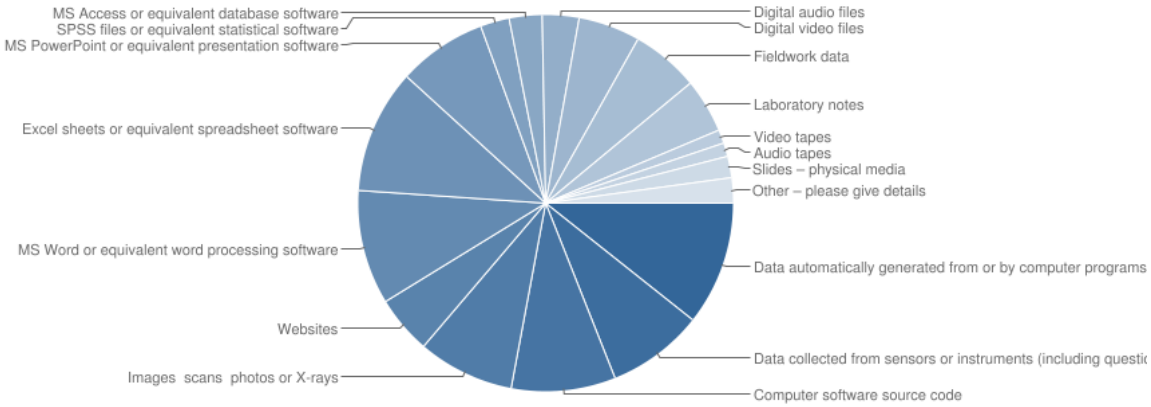

Figure 7. Data types – please tick all that apply

The spread of different data types is wide, indicating that it is difficult to focus on any one type. Interestingly, 28% is office productivity file formats (e.g. Microsoft Word, Excel and PowerPoint, or equivalent). Also, media formats account for 21% - e.g. images, audio, and video.

Respondents from Archaeology identified a number of additional data types - Physical objects (lab samples, artefacts), site drawings, ArcGIS files, bibliographic data, Illustrations based on data, current ethnographic data and 3D motion capture. ECS identified medical trial data, Mathematica [computational/modelling software]; Organisational data; calculations; rules & scoring systems; interviews and case studies.

(Some of these were identified under responses to 2.3 Characteristics of the data.)

Q2.4 Hardware/Software compatibility issues

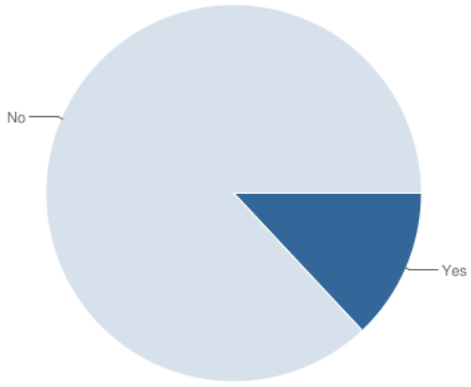

Figure 8. Do you currently have any data which is no longer compatible with existing software or on hardware media that are not now widely readable?

Only 13% of respondents to this question highlighted compatibility problems with their data.

Of these, data on floppy disks and zip disks were commonly identified. Other issues included Betacam Video tapes of interviews, paper-based data (ECS), audio cassette tape, 1930s-1950s glass lantern slides, 1950s large plate negatives, files in older versions of AutoCAD/Geoplot (Archaeology).

Q2.5 Experiences of reusing data from previous projects

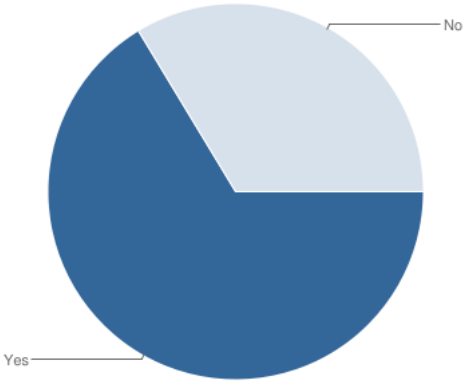

Figure 9. Have you ever re-used your own data from previous projects?

66% of respondents said they did re-use their own data from previous projects. A mixture of comments indicating this was easy, difficult or varied; with a majority (60-70%) in each school indicating reusing their own data was relatively easy.

In Archaeology there were three comments about difficulties in using old data with modern software, and also concerns about loss of data and accessibility to others. Also one respondent mentioned visiting museums to refresh his memory of objects (with varying success in locating them).

In SES difficulties in using electronic data over 10 years old and time needed to process data were identified. In ECS need for format conversion and the importance of reusing code was identified (including the use of versioning software e.g. “Forge”).

Q2.5 Have you ever used data from external sources?

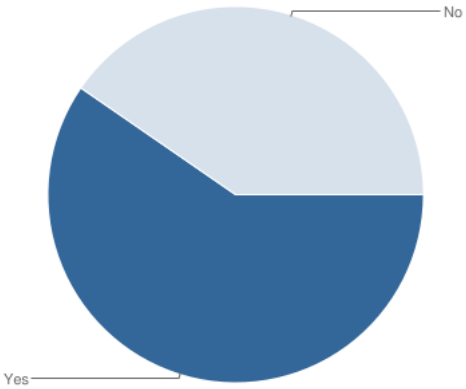

Figure 10. Have you ever used data from external sources?

59% of respondents said that they had used data from external sources.

Q2.6 Experiences of using external data

This qualitative follow-up to question 2.5 elicited interesting comments. Respondents' experiences of using external data varied widely, from being seamless, to being extremely difficult and potentially expensive.

Both ECS and Archaeology had significant usage of free data, and issues of cost and licensing inhibiting use of other data. More specific issues included licensing for software associated with specific instruments and for reproduction rights.

Archaeology also identified issues with format of the data, lack of standards in 3D data, time consuming process (and organisations slow to respond to requests for data) and good documentary/manuscript data available from Record Offices.

ECS mentioned use of open source software, difficulties in getting hold of/using and lack of support from data providers.

In SES less had used free data (availability was a problem), more had to contact the author, and sometimes to ask them questions to be able to use it.

Q2.7 Who is responsible for managing your data*

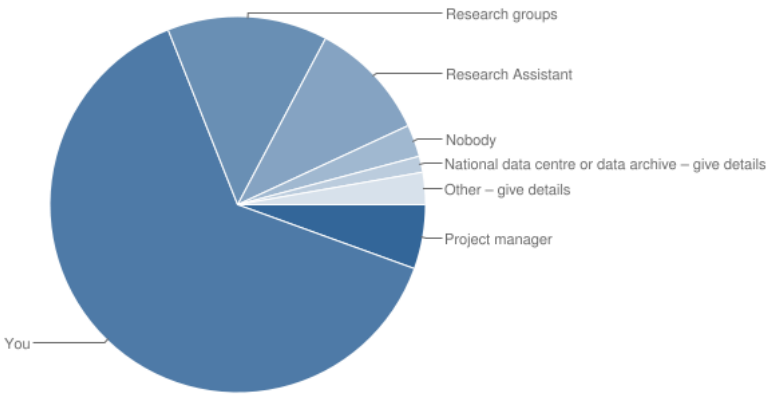

Figure 11. Who is responsible for managing your data? – please tick all that apply

There was general consensus that the individual researcher was responsible for data management – 63%. Two national data centres were named (PASCAL NoE, Archaeology Data Service). ECS and SES also identified research students and sometime researchers from another school as being responsible.

Q2.8 Where do you store your current data*

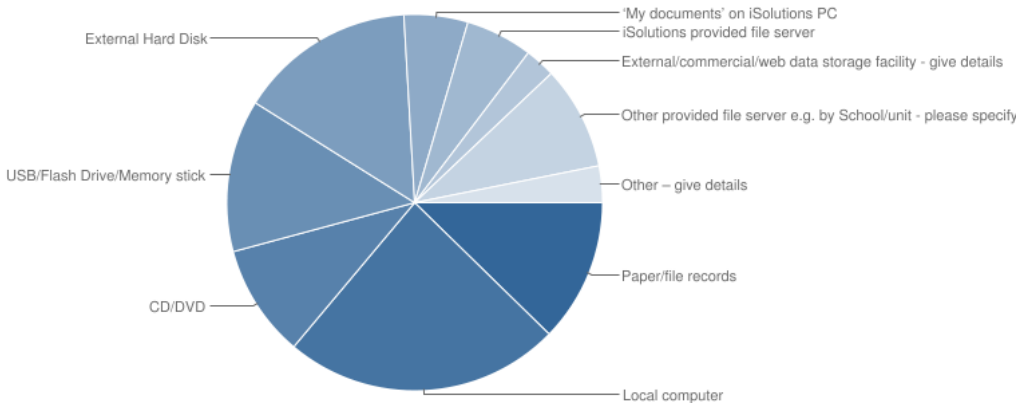

Figure 12. Where do you store your current data? – please tick all that apply

A wide array of data storage locations were highlighted in this question. 24% use their local computer, with 38% using CD/DVD, USB flash drive or an external hard disk. 23% of respondents used a file server, either at the University or off-site.

People in ECS use a number of external/commercial storage facilities including SourceForge, Google code, Google Documents, Zotero filestore and a website provided by sponsored company. Also mentioned included a remote collaborative environment server hosted by the Open University and private clouds.

There were many respondents from ECS who used file servers provided by their school. In addition they mentioned: Forge SVN server (using the Git interface), a compute server which runs both relational and RDF-based data stores, IT Innovation file servers, PASCAL Forge/EPrints/Data Repository/Video Lectures, servers bought for the project and those run by students in their group.

SES use external facilities including HECToR, National Grid Service, network drives of industrial partners and some crucial codes on Amazon's S3 service. School provided facilities included SES Research Folder, “Guide” shared drive, Spitfire Cluster (run by iSolutions), RIfI, MS home server 2TB, SESNET *fs1* file server and CT data held locally on servers bought on research grants. Some used their home PCs for backup.

Archaeology mentioned using Flickr, Integrated Archaeological Database (YAT hosted). A number of school-provided shared areas were mentioned, but appear to all refer to Resource (centrally provided by iSolutions). Note a recent audit of archaeology filestores indicates that the situation is more complex.

Q2.9 How much electronic data do you currently retain? – please tick all that apply

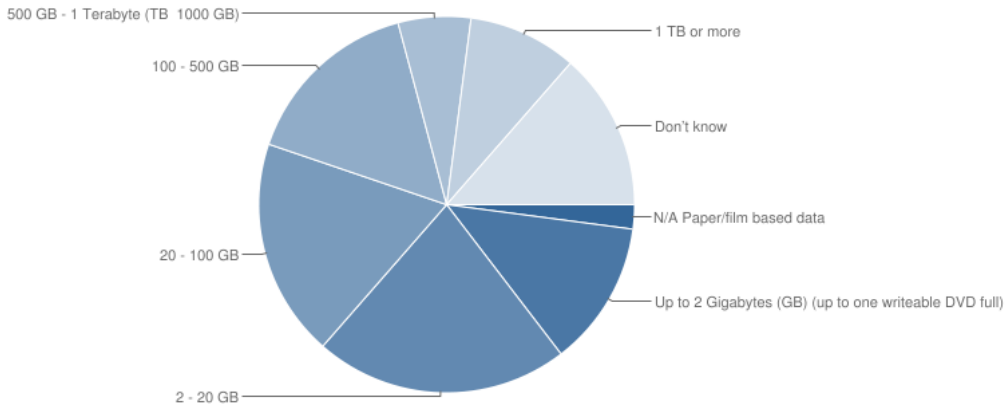

Figure 13. How much electronic data do you currently retain? – please tick all that apply – Current project

This question was split to try and ascertain data requirements for typical projects, and over the career of a researcher. We present detailed and aggregated results in Table 3 and Table 4 respectively.

It is interesting to note that for a typical project the breakdown by School indicates that in Humanities 57% of respondents held up to 100GB, with only 6.6% holding more than 100GB. Also, 30% of respondents in Humanities did not know how much data they held.

Table 3. Data stored per project, including aggregated results

Data stored	Responses	Aggregated
Up to 2 Gigabytes (GB) (or up to one writeable DVD)	13%	
2 – 20 GB	22%	
20 – 100 GB	19%	
Up to 100GB	-	53%
100 – 500 GB	16%	
500 GB – 1 Terabyte (TB 1000 GB)	6%	
1 TB or more	9%	
Above 100GB	-	31%

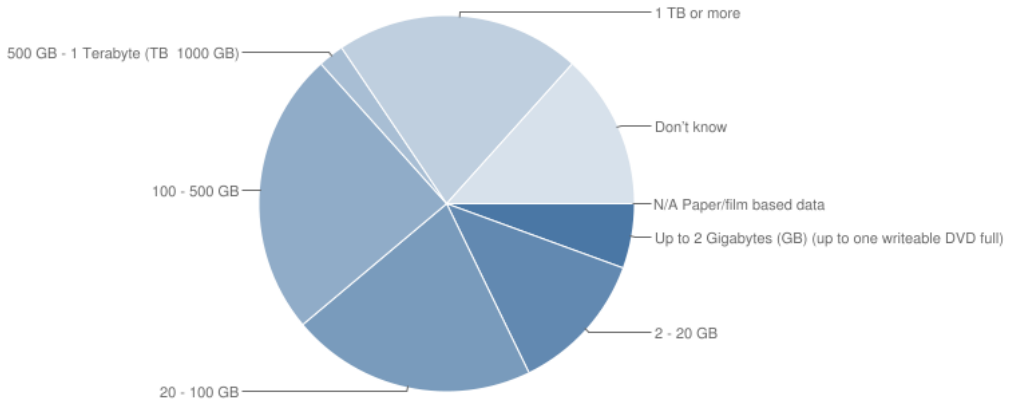

Figure 14. How much electronic data do you currently retain? – please tick all that apply – In total.

Perhaps surprisingly in the aggregated total storage for researchers, the average up to 100GB is 39%, similar to that for a typical project. More expected is that a larger fraction (47%) describes their total storage as above 100GB, with 21% above 1TB.

Here SES respondents showed a much higher percentage with over 100GB of storage requirement - 61%.

Table 4. Data stored in total per researcher, including aggregated results

Data stored	Responses	Aggregated
Up to 2 Gigabytes (GB) (or one writeable DVD)	6%	
2 – 20 GB	12%	
20 – 100 GB	21%	
Up to 100GB	-	39%
100 – 500 GB	24%	
500 GB – 1 Terabyte (TB 1000 GB)	2%	
1 TB or more	21%	
Above 100GB	-	47%

Q2.10 How long do you keep your data?

The working practices of researchers, in terms of their curation and preservation behaviour, were questioned here. The most significant result is that 42% of respondents state that they keep their data forever. This question does not include correlated data on where this data is kept, but read in line with Q2.8, it can be assumed that a significant number of users do this locally (local PC, CD/DVD, external hard drives, USB flash drives).

It shows that researchers value their research data, and therefore prefer to keep it well beyond the end of a typical project (assuming most projects last less than 10 years).

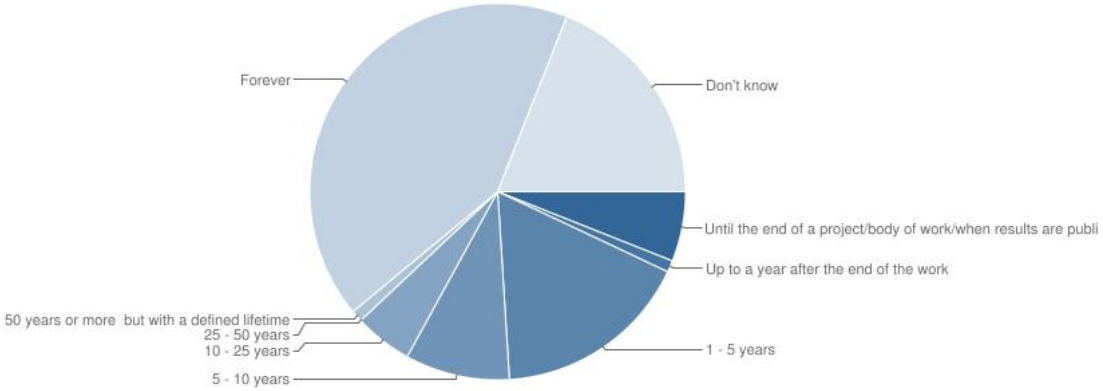

Figure 15. How long do you keep your data?

Table 5. Longevity of data storage

Time	Responses	Aggregated
Up to a year after the end of the work	1%	
1 – 5 years	17%	
5 – 10 years	9%	
Up to 10 years	-	27%
10 – 25 years	5%	
25 – 50 years	1%	
50 years or more but with a defined lifetime	6%	
Forever	42%	
Over 10 years	-	53%

Q2.11 Do you keep data for compliance reasons?

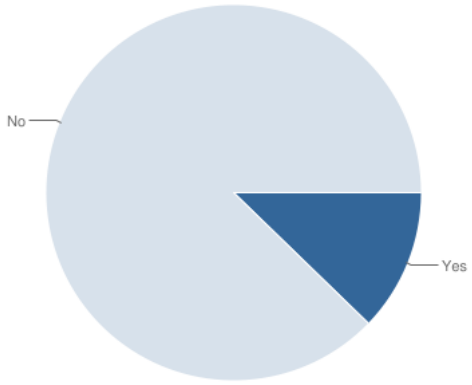

Figure 16. Do you keep data for compliance reasons?

Only 11% of respondents stated that they had to maintain data for compliance reasons.

A supplementary question was asked to ascertain how long data was held for compliance reasons. The most common length for keeping data to maintain compliance was 5 years; also other periods were given included 15 years, until the project is over, and forever. Also recycling data for other projects leading to incremental use and keeping data to 'cover our back' was mentioned.

Q2.12 Have you ever experienced storage problems due to the size of the files?

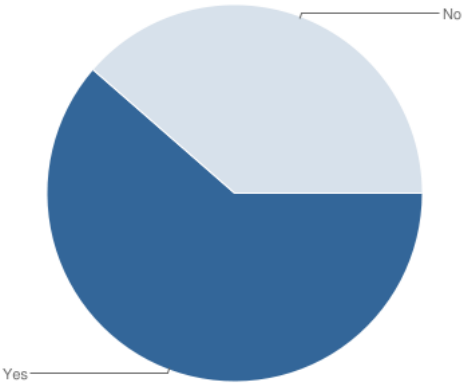

Figure 17. Have you ever experienced storage problems due to the size of the files?

The majority (61%) of users said that they had experienced storage constraints at some point, with users in SES being particularly affected (72%), compared with ECS (57%) and Humanities (56%).

Q2.12a Method to overcome storage issues*

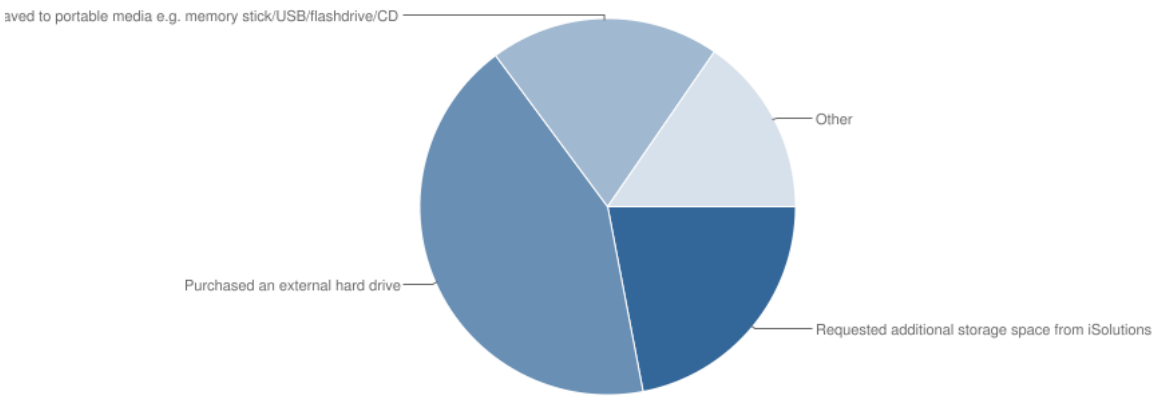

Figure 18. How did you overcome these storage issues? (please tick all that apply)

In order to overcome storage constraints 81% sought a local solution (external hard drive, CD, USB flash drive) to overcome this.

A significant number of respondents in SES (33%) and Humanities (61%) requested additional central storage from iSolutions.

ECS respondents often identified requesting additional space on ECS servers and improving capacity of existing servers. Also using external storage and removing oldest data were mentioned.

For SES there were comments identifying deleting/compressing files, an extra internal hard drive and storing data at home.

In Archaeology, school specific storage space and storing on earlier departmental computers were mentioned.

Q2.13 How frequently do you backup your data?

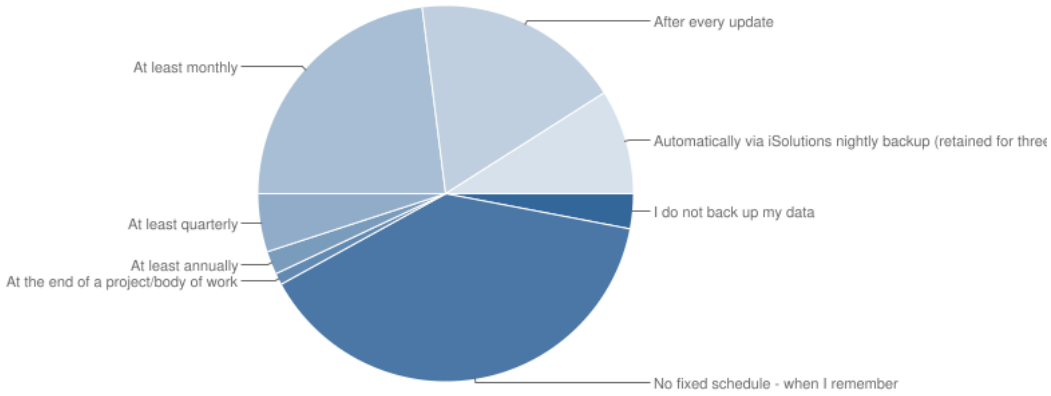

Figure 19. How frequently do you backup your data?

Backup behaviour shown in **Table 6** shows that almost half (47%) of respondents had a regular backup routine, with 7% used the University central backup system via iSolutions.

Only a very small fraction (2%) did not say they performed backups, who were all in ECS. Almost a third (32%) said that they did perform backups, but not according to a regular schedule.

Table 6. Backup frequency

Time	Responses	Aggregated
After every update	16%	
Automatically (iSolutions nightly backup)	8%	
At least annually	2%	
At least monthly	21%	
At least quarterly	5%	
Regularly, at least quarterly	-	52%
At end of project	1%	
Archival backup	1%	
No fixed schedule	32%	
Do not backup	3%	
Irregular/no backup	-	37%

Q2.14 Where do you back up your data*

This question permitted multiple responses. Over two thirds (68%) used a local solution to store a backup of their research data.

ECS mentioned using Google, public/private clouds, duplicate file servers, source code backed via ECS subversion repository, IT Innovation internal file servers (NAS and SAN based systems).

SES mentioned using local Microsoft Home Server, company backup system, emailing to themselves (web email account), home PC and Amazon S3 cloud storage. Archaeology mentioned using Integrated Archaeological Database, Windows Live Skydrive and external partners' server systems.

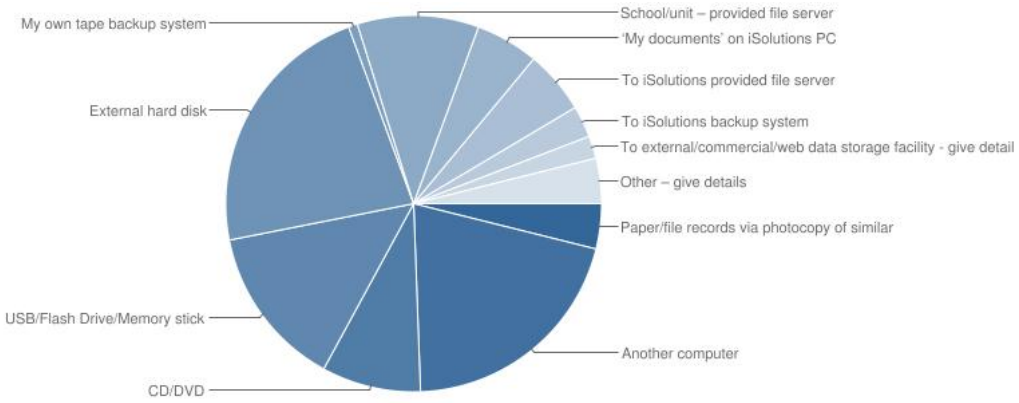

Figure 20. Where do you back up your data? – please tick all that apply

Table 7. Backup location

Backup location	Responses	Aggregated
Another computer	21%	
CD/DVD	9%	
USB/Flash Drive/Memory stick	14%	
External hard disk	23%	
My own tape backup system	1%	
Local system		68%
School/unit provided file server	11%	
"My documents" on iSolutions PC	5%	
To iSolutions provided file server	5%	
To iSolutions backup system	3%	
Server system		24%
Paper/file records via photocopy of similar	4%	
Other electronic backup	6%	
Other		10%

Q2.15 Depositing data with other services, such as the UK Data Archive

Only 10% of respondents said that they deposited data with other services. These were all in Archaeology, using Archaeology Data Service (ADS), except one ECS respondent who identified using the TPTP (Thousands of Problems for Theorem Provers) Library.

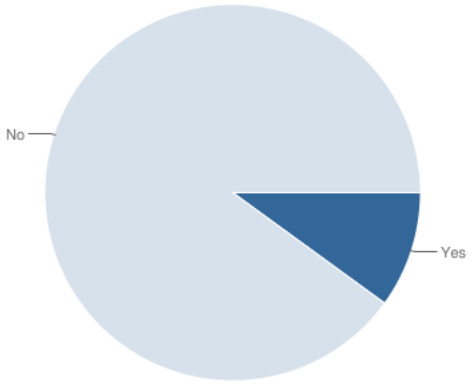

Figure 21. Do you deposit your data with other services, such as the UK Data Archive?

Q2.16 Keeping track of where data is stored, and what it relates to*

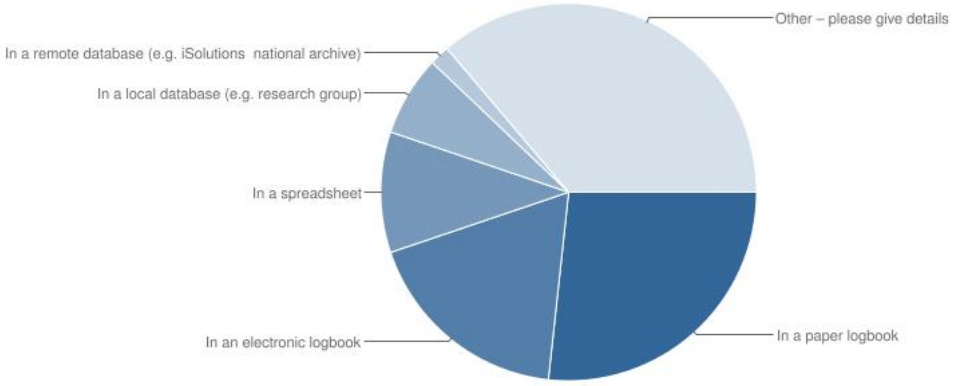

Figure 22. How do you keep track of where your data is stored, and what it relates to? – please tick all that apply

A third of respondents used a paper logbook for keeping track of their research data. Just over another third used either a spreadsheet (13%)) or electronic logbook (23%) to perform this task. Only 1% used either a local or remote database for this.

45% of researchers responded with *Other*. The most common other solutions identified were a mental record and through file/folder naming, with the following also mentioned:

- Giving each project a unique code, paper/computer files cross reference these
- Coloured pdf
- (Student) paper note books
- Electronic written reports
- Keeping track of location of paper data sheets by emailing to themselves
- Using specific programs / services (SPSS, National Grid Service Oracle Service, ArcGIS metadata file, PASCAL NoE)

There were also many comments saying their files were not well organised.

Q2.17 Allowing others access to your data – during the lifetime of a project

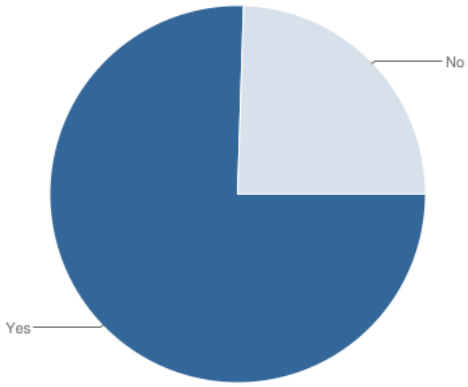

Figure 23. During the lifetime of a project, do you allow others access to data on which you are working?

Three-quarters of respondents did say that they allowed access to their data to other people during a project. The most common answer was colleagues (sometimes including students), collaborators and industrial partners. A few from all three disciplines share data to those that request it. A few in ECS host public datasets, and one in Archaeology uses the GENIE data management archive. One in SES uses their own data as teaching aids.

For those who answered *No*, reasons for this were solicited. The most common (67%) was that sharing was not required. Confidentiality or data protection issues were significant (42%), with some license agreements expressly prohibiting sharing of data (17%). A sixth of respondents also stated that that the data was not fully documented.

Other issues with sharing this data identified were - data is available as required by local research team; not wishing to give access to data they had paid for and no-one had asked for it.

Q2.18 Allowing others access to your data – after a project has ended

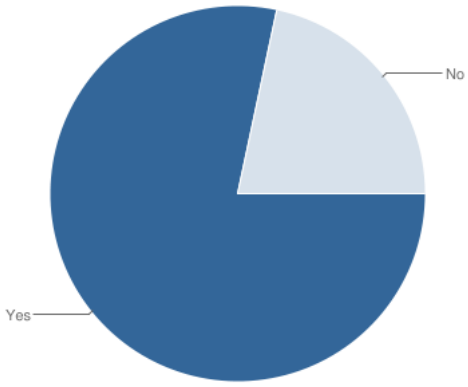

Figure 24. Do you allow others to access your data once the project is finished?

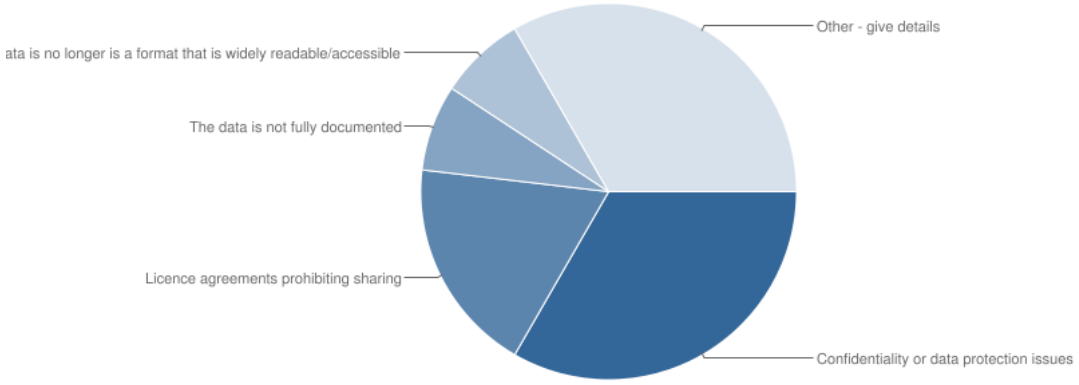

Figure 25. If no, what access issues are of concern to you? Please tick all that apply

Again, just over three quarters of respondents (78%) shared their data after the end of a project. Confidentiality and data protection was again highlighted as a significant reason for not sharing (45%), along with license agreement prohibiting sharing (25%).

Similar to Q2.17 colleagues, collaborators and partners were common answers. More willingness to share data to external users, mostly on request (where possible) – about a third in Archaeology, a fifth in ECS, but only one respondent in SES.

Other issues identified were mostly sharing not being required or no-one had asked. As in 2.17 one respondent was unwilling to share data they had paid for. One person in Archaeology indicated data was available via publications, and objects in a local museum.

Q2.19 Details of data management plans requested by funder(s)

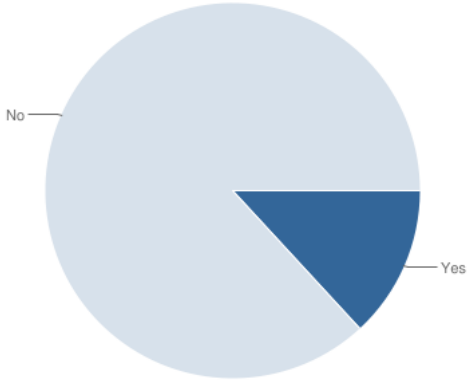

Figure 26. Have you ever been asked by a funder to produce a Data Management Plan?

Only 13% of respondents said that they had been required to submit a data management plan to their funders. This is probably biased due to the sample population – EPSRC does not require data management plans at present, while other Research Councils do (see Section 4.1). Archaeology commonly identified AHRC as requiring this, and also mentioned the British Academy and MEPF/ALSF (Marine Environment Protection Fund). SES mentioned RIfI (Research Institute for Industry - industrial consultancy unit in the school). ECS mentioned NERC and the DTC (Doctoral Training Centre).

Q2.20 Details of School policies

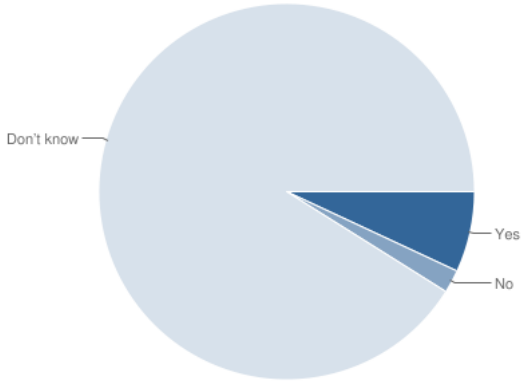

Figure 27. Are there any data preservation policies in place within your School e.g. data preservation policy, record management policy or data disposal policy?

It is clear from this question that most researchers are unaware of the existence of School data preservation policies (91%). SES mentioned the RIfI file storage policy, and the importance of storing data on a research group server to preserve IP. In ECS a data disposal policy, and informal policies in IT Innovation (e.g. two copies, two technologies, two locations) were mentioned, in addition to the Data Protection Act.

Q2.21 Would you find it useful to have university wide guidelines to manage and maintain your research data?

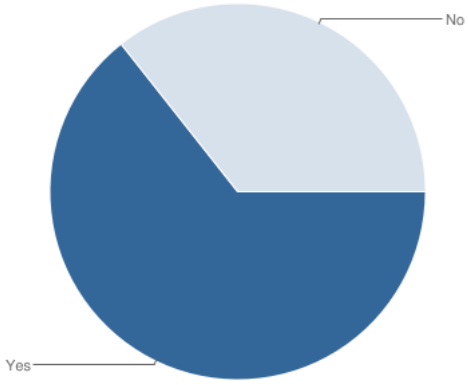

Figure 28. Would you find it useful to have university wide guidelines to manage and maintain your research data?

The majority of respondents (64%) agreed that university-wide guidelines for managing research data would be useful.

Q2.22 The biggest problem with regard to managing and storing your data

This open-ended question was aimed at finding out the most important problems facing researchers. Commonly identified issues were:

- Problems in backup due to
  - lack of space
  - time consuming (i.e. not automated)
- Organising data
  - finding and keeping track of their data (including knowing how an image/spectra/etc. was obtained),
  - issues with version control especially for code.
- Lack of space on fileservers (mostly an issue for SES/Archaeology),
  - problems caused by large files (storage, processing). Hadoop cluster + hdfs storage was identified as a solution for the latter by one person.

ECS also identified the need for better guidance for good practice and issues with remote access, not all data being publishable, developing bespoke systems to manage data, limits on data crawling in public data sources (delegate a crawl across numerous IP addresses to get round this) and not using iSolutions for support.

Archaeology also identified a need to make data accessible after the project is finished.

Q2.23 How can the University make data management and storage easier for you

Question 2.23 aimed to solicit suggestions on how the institution can assist researchers with their data management. The main issues identified were:

- More guidance needed (but not rigidly imposed rules – diverse needs of researchers) – including what facilities/services were available, data management training
- A need for more automated backup, especially for data on desktop/laptops
- More storage space on network drives (20/100 GB/unlimited) and better ability to access remotely. Comments about lack of space for PhD students, enterprise workers and having a high capacity my documents/known archive space for each person
- Some in all three schools requested an '*EPrints for data*'

ECS respondents also identified a need for a *Git* versioning server (with enough hard disk space), software to help with data management/storage and a service to scan lab note books.

Archaeology identified needs for uploading data to the University while in the field, developing an ethnographic/archaeological archive, more space for physical data ability to archive data to the ADS.

SES also identified problems with accessing My Documents and using memory sticks on library computers, and need for a system to zip files and archive logically (Linux), providing external hard drives and initial advice on data management.

2.2.1.3 Further Comments

Q3.1 Any further comments

The questionnaire concluded with an open question to capture respondents other thoughts. These are broken down by School below.

Archaeology

- Happy with current set-up - large amount of backed-up space (J drive)
- Would be helpful to have a reference to query on data ownership/advice on sharing primary data
- I keep records of meetings etc rather than actual research data
- Space to store paper data is an issue (and the money or time to photocopy to keep backups). We have taken to taking photographs of workbooks in the field so that we have some electronic back-up of the day's work.
- The main issue for me is about deposition and long-term curation of (digital) photo and video archive
- What to do with paper archives of FRU (Faunal Remains Unit)?

Electronic & Computer Science

- Large server needed for storing video data
- Run training session "how to manage your data resource"
- Issues in preserving AV data with complex rights issues limiting retention/use - completed DRAMBORA risk assessment for PrestoPRIME project (threats including storage systems, file formal/bit-level preservation e.g. migration) - including total cost of ownership over 10 years
- Main issue for old data where author has left the university, need to avoid introducing more admin
- Avoid a complex centralised bureaucracy - just lots of storage with good backup, everything else is optional
- Questions in survey pitched more to empirical data researchers thus harder to answer for modellers

Engineering Sciences

- Suggestion to provide all PhD students with 1-2 TB external hard drive
- Need for 4 TB of backed up space

2.2.2 Interviews

Interviews have been conducted with some of those who filled in the questionnaires. To date, this is as follows:

School	Academics & Research staff	Graduate students	Total
Archaeology	12	18	**30**
Electronics and Computer Science	8	4	**12**
Engineering Sciences	7	1	**8**
Total	**26**	**21**	**50**

The interviews were semi-structured, as shown in Appendix II. The interviews were split into three major sections:

- Managing data
- School Policy and data management planning
- Collaboration and sharing

Interviews were one hour long, with two interviewers, one subject specialist and one librarian. In each case detailed notes were taken, along with an audio recording. Due to resource limitations transcriptions were not produced, but notes and audio recording cross-references in order to summarise our findings.

Despite the different subject areas and types of user, many common themes emerged and differences in responses were minimal. This data complements the questionnaire responses to give us a picture of researchers' data management experiences. Together these two datasets provide us with the foundation for future planning.

2.2.2.1 Managing Data – Storage

A significant issue for researchers was the availability of adequate, secure storage that was easy to use. This is seen as one of the most pressing issues at all levels, from senior researcher to PhD students. While availability of storage at School level varies, as indicated in the questionnaire results and data infrastructure audit in Section 0, the interview responses share many common themes.

- By far the majority of users did not use the University storage facilities under *My Documents* or '*J*' drive. Instead data was either stored on personal pc or laptop. Backups were most often done using purchased hard drives (many research students buying them out of their own pocket) kept at home, and memory sticks, discs or in emails. When asked why they did not use University storage, this was either because they were not aware they could request more storage, they were worried that they would lose control of their data (especially if they were to move on), ease of access off campus or worry that it would get "lost" in the system. Those who had experienced loss of data through not backing up, not surprisingly over-compensated by having a variety of back-ups, often up to 6 external hard drives and thus created a new problem of ensuring that all versions were the same.

- Despite the fact that iSolutions offer to increase quotas on request, many did not realise this or thought their need would still be greater. Many store huge amounts of high speed imaging data or graphics and small increases nowhere near enough.

- In Archaeology and SES there are huge amounts of physical data (models, artefacts, pottery, bones for example) kept in offices or in labs. Some of these are labelled and organised, but many are not – simply being kept in bags/boxes in the office. Also a lot of paper files, notes and photographs which could not be backed up.
- Most people seem to keep data indefinitely as never sure if it might be needed again or if others may want to use it. With rapidly developing software and technology, this has meant that some have data is in formats which cannot be used. In some cases this is not recoverable as the disk has deteriorated and is unreadable. iSolutions have helped in some cases, but it is an ongoing problem for others.
- Limited use of Data archives such as ADS and UK Data Archive, sometimes it is felt that data may not be important enough.
- Legacy systems: computers on 'SESNet' store My Documents locally on PC hard drive. It is neither backed up or available on public machines, and no access to Networked drives.
- Iridis (high performance computing) – offers limited space and have been issues transferring data to/from this. Also, if people are encouraged to move low-use data, then more guidance is needed to using alternative locations (one valuable dataset ended up on lots of external hard drives).
- Deleting data hardly ever happens (especially true in Archaeology) as cannot be reproduced in many cases, especially if dependent on a particular moment in time or the environment.
- Not sure how long iSolutions keep things for.
- Shared storage space may mean that no-one takes responsibility, especially if the objects stored are models, objects, etc.
- Papers and objects vulnerable to damage and wear and tear over time.

2.2.2.2Managing Data – Access

Here interviewees were asked to describe how they keep track of and are able to access their research data.

- Most people have a system of storing data which works for them using a folder structure with appropriate metadata. However, metadata can be very subject specific and personal – most people agreed that others would find it difficult to know where to look or make sense of it without extra information such as personal knowledge or a spreadsheet.
- Need for better applications for secure sharing of large files with collaborators – also need for 'Download Manager' to avoid file corruption. Some interest in SharePoint. Code logging software (not available from iSolutions) would be useful when multiple people are working on a programme.
- Confidence in access to personally stored data gives way to doubt when asked if others could easily find their way around the data. General acknowledgement in principle that a template with metadata might be a good idea, but worry that to cover so many different subjects the metadata would have to be so general as to be useless.
- Could be problems when a PhD student or post doc leaves in making sure that the supervisor has access to their physical and electronic data and that it is labelled in a way they understand. Would be good to have a consistent system in place.

- Concern that access to data will no longer be possible when students have left the University and free software no longer available.
- Some students have trouble getting VPN to work so buy their own external hard drives.

2.2.2.3Managing Data – Compatibility

Some users had issues with compatibility of data with software availability over time. This was highlighted in the questionnaire and probed further here.

- Has been a problem for many people – significantly senior staff who have been working with data for a long time. Software can change without warning and they have not got the time or technical ability to keep up to date with new formats.
- Computers can be linked to instruments/testing machines which are old, maybe as much as 10 years, and are difficult or impossible to replace without replacing the instrument as well. Older operating systems can mean problems accessing the network and installing software.
- A case of someone reverse-engineering some data – trying to solve own compatibility issues because mainly images.
- Have used open source software in some cases but others mostly do not use it so data has to be constantly re-formatted.
- 3D data cannot be saved into a good useable format.

2.2.2.4Policy, Guidance and Training

The issue of support for researchers, in terms of policy, guidance, and training was a major theme within the interviews across the spectrum of interviewees.

- There was a general lack of knowledge, particularly amongst the Ph.D students, as to what was available both in terms of what iSolutions could offer, storage and who to contact for help.
- Many students asked for training to be given in data management – early in their studies and to be given a small project with which to practise the knowledge so it would not be forgotten easily.
- Academics also acknowledged that this was an area where students needed skills training. Without experience and training, students did not realise the need for a system until too late into their research and ended up with data in lots of different places which needed to be pulled together.
- Both academics and researchers tended to use their own systems but most thought that University guidelines would be useful as long as they were flexible enough to cover all subject areas meaningfully.
- Although some people had worked out automated systems to back up data on external hard drives, others had not and would find guidance here useful.
- Trade-off seen between access and storage advantages and the inconvenience involved in having to do things in a certain way – especially if mid-experiment.
- Happy to follow if professionally done and not too general.

2.2.2.5Data Management plans

Within this sample group, only the Humanities researchers are generally required to complete data management plans (see Section 4.1), hence the relative lack of interest in these questions.

- Most students had never made a plan nor had heard of it. Some academics had done one but they were in the minority. In SES, there were few formal plans and those had usually been at the request of industrial funders. There had been some informal planning – storage space needed and how long it would take to run the information. Most of the issues for SES focused on data security and confidentiality.
- Some people work with commercial data which has different requirements to academic research and so there can be many approaches as companies have their own specific archiving requirements. Would be better if there was just one way.
- Felt it would be a good idea for PhD students to do one.

2.2.2.6Collaboration/Sharing

The idea of sharing data, both internally and externally to the University, was probed and elicited a variety of responses. These largely backed up the quantitative data from the questionnaires.

- Within Archaeology, there were no real issues with sharing data. Almost everyone was happy to share, especially when the project was finished. Reservations were either because data was not complete and could be misinterpreted, or uncertainty as to ability to share as some objects had been borrowed by special agreement with Museums, etc and it was felt that the data may not be theirs to share.
- In SES, there seemed to be much more reluctance to share due to: worry that the data would be used in ways other than was intended, worry that data may not be understood correctly, easier to create own data, and unwillingness to trust people who you may not know very well that the data would be correctly acknowledged and cited.
- ECS are mostly happy to share. However, they share the same issues with SES concerning trust outside the University as they also have had shared data which has been unacknowledged.
- Issues of confidentiality, data protection, etc. – so outside personal control.
- If ownership of the data is uncertain or felt it belongs elsewhere, then there is reluctance to share for ethical reasons.
- Felt that drawings should be freely available. Some re-draw images to avoid expensive copyright fees but this means that each copy becomes increasingly inaccurate and is a waste of time.
- Sometimes this has only been possible because of networking and knowing the right people to ask favours from. This can be difficult for new researchers, though the trust element comes into play here.
- Not enough knowledge of copyright issues to know whether this is always possible.
- Long term security and easy access.
- Some things too personal to share – fieldwork log books for example.
- Not sure who has the responsibility for a borrowed object – University, person borrowing – able to share?

- Feel there is zero connectivity between institutions at the moment.

2.2.2.7Wishlist

At the end of each session, interviewees were additionally asked for a number of wishes to help with their research data management. This was similar to the approach used in the kick-off workshop, described in Section 3.

- An easy-to-use Data Management plan which is flexible enough to be used by an entire institution but also takes specific subject needs into account. An ability to link to subject-specific directories and taxonomies would be useful.
- Training to be delivered to all new PhD students and researchers, and academics if desired. Who is to deliver the training? Training might be more a way of looking at the issues, a way of thinking, rather than specific systems which may be replaced.
- Manpower to support and update any system.
- A dedicated archive keeper for non-digital records.
- Ability to locate material more easily in an institutional system – present impossibility of finding anything on University web pages and in file structure.
- Better ways and facilities of sharing documents between agreed people.
- To be warned when software and existing technology is about to expire and told what to do about it. Would also be useful to have a "Library" of old programmes which could be used to read old data and convert it.
- A way to archive emails for posterity as an academic record – in the way that letters were archived in the past.
- Guidelines as to how long official reports or data should be kept – useful for staff working with admin files as well.
- An extremely large networked drive with the option to read other peoples' drives and vice versa.
- More storage space with larger quotas given without having to keep requesting. Many students from other cultures do not feel they can ask for more so buy their own hard drives instead. It was suggested that 100 gigs per PhD student would be a better allocation. Also that it be easier to request more space instead of emailing for each little bit.
- More knowledge as to rules of ownership.
- Data management as a data store – access to an unrestricted volume of date where all information can be deposited, maintained, backed up, time logged and easy to access.
- Fire and hard disk crashes left the feeling that a secure centralised form of storage would be best as long as could log in and manipulate data in own way.
- Front end applications allowing connections to be made via Web 2.0

- Ability to connect with others using same interests and literary sources which could be picked up from raw data, literary data (eg.library records) and other global projects. In the way that Amazon suggests other books from your past purchases.
- A “toolbox” of University help and guidelines so you could select what you needed.

2.2.3 AIDA

In order to provide a top-down view of research data management at the University of Southampton, we have chosen to adopt, and adapt, the Assessing Institutional Data Assets (AIDA) toolkit[10]. This is being carried out as part of the Integrated Data Management Planning Toolkit and Support project[11].

The AIDA toolkit was designed principally to assess digital asset management capability of institutions, and was targeted initially at digital preservation. Hence the audience was records managers, librarians, data curators and repository managers. The context of the IDMB is different, however, as it includes the earlier stages of data management, from creation, through usage, ending up with preservation and curation.

The IDMB team has worked with Ed Pinsent, at the University of London Computer Centre, to adapt AIDA for our own use at a *departmental* level. AIDA is a self-assessment toolkit and comprises of questions in three major categories, following the *three-legged stool* digital preservation model developed by Cornell University:

- Organisation
    - Ownership and management
    - Policies and procedures
    - Policy review
    - Sharing of Research Data / Access to Research Data
    - Preservation and continuity of research data
    - Internal audit of research activities
    - Monitoring and feedback of publication
    - Metadata management
    - Legal compliance
    - IPR and rights management
    - Disaster planning
- Technology
    - IT environment and infrastructure
    - Appropriate technologies in place
    - Ensuring availability and integrity
    - Integrity of information
    - Obsolescence, format management
    - Security of environment
    - Metadata creation
    - Institutional repository management

Resources

    - Financial sustainability plan
    - Resource allocation
    - Risk analysis
    - Sustainability of funding
    - Staff skills
    - Staff numbers
    - Staff training

The detailed questions for each category are shown in Table 8, Table 9, and Table 10.

[10] http://aida.jiscinvolve.org/wp/

[11] http://www.jisc.ac.uk/whatwedo/programmes/mrd/supportprojects/idmpsupport.aspx

Use of the tool at Department/School level has been attempted, although accurate results are difficult to obtain. Each question requires an answer to estimate the level of confidence in capability for each topic. This is difficult at the School level when there are a large number of researchers. In the case of SES this is in the hundreds, and so it is difficult to make this assessment. Here we have made best efforts to estimate the level of activity, and there is therefore a significant degree of uncertainty in the results. On a five-point scale we estimate error is at least one point either side, i.e. +/-20%, although errors could easily be larger. Due to the level of uncertainty we have not presented absolute numerical scores, but instead present a brief narrative assessment that describes the level of activity and capability within the bounds of uncertainty. It is, however, a useful exercise in order to identify areas for improvement, and it is in this light that AIDA has been used here.

The forms have been completed in a collaborative way between relevant members of the IDMB team in each School, in conjunction with senior managers. Note that the University is reaching the end of a consolidation exercise for its IT provision. Therefore research data systems and their management is currently moving from local to institutional level of responsibility over the space of this project. The AIDA results try to capture this transition phase, with an eye to the equilibrium future state.

2.2.3.1 School of Chemistry

The School of Chemistry is in a period of structural transition – this report considers the period up to 1st August 2010 and subsequent reports will highlight the new structure. This report therefore provides a baseline, against which the new structure will be compared.

The School of Chemistry comprises over forty research groups, along the following themes:

- Synthesis & Catalysis
- Chemical Biology and Electrochemistry
- Interfaces & Materials

This structure, implemented in 2004, was a departure from the traditional structure of Chemistry departments based around the sub-disciplines of Inorganic, Organic and Physical chemistry. The School has a tripartite mission – education (UG & PG), research and commercialisation, each bringing different needs with respect to data sourcing and management.

Funding comes from a variety of sources, including Research Councils, charities, the European Community, US Government sources and Industry in the UK, Europe and the USA. Researchers carry out extensive experimental, computational (including HPC) and theoretical work.

The core activities of the School generate large volumes of raw and processed data arising from synthesis, characterisation, modelling and computation captured by traditional lab notebooks weakly associated with digital storage. The scale is highly variable with multiple and conflicting formats and the extent of metadata potentially complex. The School houses a number of notable support facilities, including:

- UK National Crystallography Service
- NMR facility
- Mass spectrometry laboratory
- X-Ray diffraction facilities

which, as such, demonstrate greater concern with users data and its curation than the average research group.

There is a highly observed and well developed safety culture with the result that all information relating to safety is highly organised, quality controlled and curated across the School. Other types of research input and output are much more variable across research groups.

The following sections summarise results of the AIDA assessment, however it should be noted that these are 'averaged' responses, due to the difference between safety and other data outlined above.

2.2.3.1.1 Organisation

Ownership and management of research data appears to be varied. A solid research culture and best practice is exercised across the School, including monitoring and feedback for publications and data publication. The latter is encouraged and actively pursued in some cases. Knowledge of data responsibilities and policies appears unified within the School. Sharing, access, preservation and continuity of data is consolidated, with some pockets of excellence. Metadata management and sharing rights management are both areas where improvements could be made. The School takes care to ensure legal compliance and safeguarding IPR and this is embedded within the culture. There is, however, room for improving awareness of relevant legislation.

2.2.3.1.2 Technology

Historically the technology infrastructure has been localised at the School level, although this is now being consolidated with the institutional IT provision (iSolutions). Due to this situation, appropriate technologies, availability, data security and managing obsolescence of data formats has been performed locally. There is a gradual move to more institutional level of consolidation in areas such as metadata management, and support through our institutional (EPrints) repository.

2.2.3.1.3 Resources

The availability and allocation of resources for data management is localised, with an acknowledgement that staff support could be significantly improved. There is recognition that a financial basis for this is required, and there is some consolidation of how income might be generated from research data in this context.

2.2.3.2School of Engineering Sciences (SES)

The School of Engineering Sciences performs research across a wide-range of areas, with groups in:

- Aerodynamics & flight mechanics;
- Astronautics;
- Bioengineering;
- Computational engineering design;
- Electro-mechanical engineering;
- Energy technology;
- Engineering materials and surface engineering;
- Fluid-structure interaction;
- National Centre for Advanced Tribology (nCATS).

Its researchers carry out experimental, computational and theoretical work and they have access to numerous laboratory and high performance computing facilities. The School carries out work that is funded from a variety of sources, including research councils, the EU and industry. It houses a number of research centres and partnerships with industry:

- Lloyd's Register University Technology Centre (LR UTC) in Hydrodynamics, Hydroelasticity and Mechanics of Composites;
- Royal National Lifeboat Institution Advanced Technology Partnership on Maritime Engineering and Safety (RNLI ATP);

- Ministry of Defence/Lloyd's Register Centre of Excellence for Marine Structures;
- DePuy International University Technology Partnership (UTP) in Bioengineering Science;
- Rolls Royce University Technology Centre (UTC) for Computational Engineering
- Airbus Noise Technology Centre;
- Microsoft Institute for High Performance Computing;
- National Centre for Advanced Tribology at Southampton (nCATS).

2.2.3.2.1 Organisation

In this School the research group is the effective operational unit for much of the data management activity, in the context of project teams which may span research groups; and include members from outside the School. This is the case for ownership, policy, procedures, sharing, access, preservation and continuity of research, metadata management. A culture of good research practice, including monitoring and feedback of publications, legal compliance, IPR and rights management, is embedded across the School.

2.2.3.2.2 Technology

Technological infrastructure is being consolidated with the institutional IT provision (iSolutions), and varies across the School with some localised provision currently. Appropriate technologies are deployed at a localised (project) level to meet specific needs, within the background of core provision. Institutional repository support is localised but has recently been provided (via EPrints) but has yet to be used extensively. Areas which are localised and in which there is room for improvement include: availability, information integrity, dealing with obsolescence, and metadata management.

2.2.3.2.3 Resources

Resources tend to be managed at the project level, notably technology, although some financial sustainability planning is done at the School level. Staff skills and development is generally consolidated, although staff numbers are limited.

2.2.3.3School of Humanities (Archaeology)

The School of Humanities carries out research in a number of related disciplines:

- Archaeology;
- English;
- Film;
- History;
- Modern languages;
- Music;
- Philosophy;

It also houses a number of research centres:

- Centre for Medieval and Renaissance Culture;
- Centre for Transnational Studies;
- Southampton Centre for Eighteenth Century Studies;
- Parkes Institute for the study of Jewish/non-Jewish relations;
- Centre for Applied Language Research.
- Centre for Maritime Archaeology
- Archaeological Computing Research Group
- Centre for the Archaeology of Human Origins

Humanities also engages in consulting activity, primarily in the area of archaeological geophysics under the aegis of Archaeological Prospection Services of Southampton. In this study we have

focussed on researchers in archaeology, as they are also one of the pilot study groups. Where appropriate Faculty level information has been incorporated. We anticipate that the completed IDMB project will be able to draw on a full AIDA appraisal of the Faculty of Humanities in order to contextualise data management policy and practise within the archaeology pilot.

2.2.3.3.1 Organisation

Ownership and management of data tends to be localised, and carried out at a research group level with project teams managing data explicitly. Within this context, policies, procedures, preservation, audit, monitoring and feedback on publications, IPR and metadata management are all areas for improvement.

Sharing and access to research data is well understood and largely consolidated, with pockets of excellence for sharing beyond the institutional boundary.

2.2.3.3.2 Technology

Technological infrastructure is largely localised but is being consolidated with the institutional IT provision (iSolutions). This includes use of appropriate technologies, availability, integrity of information, managing obsolescence, an dmetadata creation. Within this context hardware and software provision is consolidated, and institutional repository support is becoming more so.

2.2.3.3.3 Resources

The management, allocation, and financial sustainability planning for resources is carried out at a local level. Technological resources tend to be managed at a project level, and is moving to an institutional level, within the context of our IT service restructuring. Staff skills and development is also localised, with the limited number of staff available.

Table 8. Departmental AIDA questions – Organisation

ORG 01: Ownership and Management	Our Department has a formal statement on ownership and management of research data
	We know who owns data and who is responsible for managing it
	We know who owns subsidiary documentation and notebooks
	We have written guidelines to support the formal statement
	The statement is shared in the Department
	The statement has been accepted by the Institution
ORG 02: Policies and procedures	We have written policies and procedures for management of research data (e.g. a Data Management Plan)
	We have written policies and procedures for preservation of research data
	We have written policies and procedures for data sharing
	We know how we will manage our research data now and in the future
	We have written guidelines to interpret the policies
	The policies and procedures have been implemented at the highest level
	The policies and procedures are followed
	The policies and procedures are fully integrated with each other
	The policies and procedures relate to our research data in a meaningful way
	We include statements about data sharing when we apply for a research grant
ORG 03: Policy Review	Our written policies and procedures are subject to regular internal review
	Our written policies and procedures are assessed by a Committee / Panel / Reviewers
	We take action, amending and revising the policies after review has taken place
ORG 04: Sharing of Research Data / Access to Research Data	We can, and do, share our research data as appropriate.
	We share research data with our immediate colleagues
	We share with others in the University
	We share with others outside the University
	We re-use and re-purpose data (secondary use is allowed)
	We collaborate with each other and with others in the scientific community
	We have full access to our data
	We allow access during the project lifetime
	We allow access after the project lifetime
	We share data in a timely way
	We share data with as few restrictions as possible
	Our sharing is licensed as needed

ORG 05: Preservation and Continuity of Research Data	We are aware of the need for long-term continued availability of research data We have made successor arrangements for research data (when members of staff leave) We know about our preservation requirements We know how long the research data should be retained after its immediate use has ended, or on completion of the research We have a contingency plan We use external services for deposit / preservation We are confident that our research data is protected
ORG 06: Internal Audit of research activities	Researchers keep track of their actions Researchers keep formal records The records are updated regularly Notebooks, spreadsheets, and databases are used Researchers record the stages in the creation and usage of their research data Changes to the data are recorded Because of these activities, we know what researchers are doing with their data
ORG 07: Monitoring and feedback of publication	We publish our research data as appropriate Our research data is citeable We have knowledge of the use being made of our data We monitor on a regular basis the usage that is being made of our research data We follow timescales for release of data after publication We ensure that our data sources are acknowledged
ORG 08: Metadata management	We use metadata schemas to annotate and organise our research data We put some or all of our research data in a repository that requires the supply of metadata We are aware of external metadata standards and use them Our data is fully retrievable Our data can be discovered easily Our research data is self-documenting Our use of metadata is supported by Information Management professionals (for example a repository manager, digital librarian or archivist) We create quality metadata (provenance, context etc) which makes the research data understandable to secondary users

ORG 09: Legal compliance	We retain our research data in line with legal compliance reasons We ensure correct and legal usage of the research data during lifetime of the project Where appropriate, we protect the confidentiality of the research data during lifetime of the project We lock down research data after the project completes (e.g. for legal proof or protection of patents) Access to research data (by staff, internal or external users) is managed and monitored at all times We are aware of legislation that affects our research data (e.g. Data Protection, Freedom of Information) We have procedures to ensure ethical use is made of our research data
ORG 10: Intellectual Property Rights and rights management	We have clarity on the ownership and rights associated with all of our research data Researchers understand their responsibilities for rights management We make use of Creative Commons and Scientific Commons to allow the correct degree of sharing and protection Attribution is clear and well-managed Mechanisms are in place for the automatic detection of rights expiry, where needed Mechanisms are in place for managing access to our research data in line with IPR, copyright, attribution etc. We are confident we can safeguard IPR, proprietary data and patentable data
ORG 11: Disaster planning and continuity of research	We have formal arrangements in place to ensure research could continue in case of data loss We have a written disaster plan We have a written plan for continuity of research Our plans are tested, reviewed and updated regularly Our plans are communicated to all the staff involved Our plans have organisational acceptance The disaster plan has an owner and project manager

Table 9. Departmental AIDA questions – Technology

TECH 01: Technological infrastructure	We have an infrastructure appropriate for our research data management needs The infrastructure supports the amount of data we hold Our departmental set-up is harmonised with the central IT infrastructure We have a formal list of our IT assets Our departmental set-up allows us to share research data as needed Our departmental set-up is supported by appropriate SLAs Infrastructure investment is planned to meet Departmental needs
TECH 02: Appropriate technologies	We have the correct sort of software for research data and its management We have the correct sort of hardware for research data and its management The software and hardware matches the anticipated lifespan of the data The software and hardware is appropriate for storage of our research data The software and hardware is appropriate to allow Departmental access to our research data The software and hardware allows us to share our research data (e.g. through deposition in public databases)
TECH 03: Ensuring availability and integrity	We know where our data is backed up; numbers and locations of all copies are known Our data is backed up with a reliable frequently Backup allows for anticipated growth of our research data collection Multiple copies of research data are synched We do not rely on local copies (including local drives, laptops, memory sticks etc) for storage We discourage "offline working" in favour of working with centralised and managed storage We have sufficient network space for storage and the file sizes of our data present no problems Data storage program includes offsite storage and/or outsourced external storage We have arrangements for storing non-digital data, including paper We deposit our data with other services

TECH 04: Integrity of information	We have mechanisms to detect data corruption We have mechanisms to avoid data loss We have mechanisms to repair damaged or corrupted data Preventive detection checks take place regularly We have a media testing program for CDs and DVDs We consider our research data to be safe from corruption Data that is released for sharing is validated and verified in line with accepted best practice and is of high quality
TECH 05: Obsolescence	Our research data is kept in file formats that will support its longevity; we work to standards for data formats We do not use storage media (like CDs, DVDs) for long-term storage Research data is not kept on local drives, laptops or memory sticks We have a good understanding of obsolescence issues Obsolescence is dealt with pro-actively We have a file format registry
TECH 06: Changes to critical processes	Not needed
TECH 07: Security of environment	The hardware on which our research data is kept is secure Our working environment is secure Information environment is analysed systematically Research data is stored in an access-controlled area External threats and denial of service attacks are addressed by regular analysis
TECH 08: Security mechanisms	Not needed
TECH 09: Implementation of disaster recovery plan	Not needed
TECH 10: Metadata creation	We make use of automated metadata tools to create and manage metadata as appropriate Tools are useable and are used Tools allow us to locate and use the research data Tools are integrated with the research data lifecycle
TECH 11: Institutional Repository	We use a repository to manage and store some or all of our research data The repository is usable and is used Research data is protected in the repository The repository has appropriate security embargos in place The repository allows sharing of research data internally The repository allows sharing of research data externally

Table 10. Departmental AIDA questions – Resources

RES 01: Financial sustainability plan	**The business plan supports the sustainability of our research data** **We generate income through our research data** **Our research data collection is self-supporting**
RES 02: Review of business plan	
RES 03: Technological resources allocation	Sufficient money is being invested on the technology we need for our research data (not just storage) There are dedicated funds available for technology development in support of our research data Technology Watch is in place for emerging technologies Future technological requirements are anticipated Department is capable of assigning the necessary technological resources to the research data collection
RES 04: Risk analysis	We have a formal risk management plan in case of data loss Risk analysis is based on existing standards
RES 05: Transparency and auditability	
RES 06: Sustainability of funding for research data	There will be enough money to keep our research data safe Funding is inbuilt to the core function of our Department
RES 07: Staff skills	Department has the requisite skills available to manage its research data Our funding enables the steady maintenance of core staff skills
RES 08: Staff numbers	Department has enough staff to manage its data
RES 09: Staff development	Staff are competent in research data management Staff skillsets have currency Staff skillsets evolve in line with technological changes We have a professional development and training policy We have a training budget

Table 11. Departmental AIDA Stages

Stage 1	**ACKNOWLEDGE / NO ACTION**	Very low confidence. Nobody in the Department is doing this. We don't have any formalised financial or staffing policies. We do not have financial autonomy. We have no evidence of any action. These things are implied rather than actually carried out. We don't meet the benchmark but we acknowledge this is the case.
Stage 2	**ACT / LOCALISED**	Some confidence. At least one person in the Department is doing this. We have evidence of some local activity. Practices can vary, and are ad-hoc and inconsistent. Work on this is still unfinished or it has only just started. Financial allocation in this area is uneven.
Stage 3	**CONSOLIDATE / CO-OPERATE**	Medium confidence. At least three people in the Department are doing this, and are doing so in harmony with each other. Practices are consistent within the Department. Their activities still cover only a few defined areas of managing research data, not everything. These actions are local and only affect our Department. We are not yet harmonised with the entire Institution.
Stage 4	**UNIFY DEPARTMENT / INTERNAL INTEGRATION**	High confidence. Everyone in our Department / Research Group is doing this. These actions are fully in place. All defined areas of financial practice, funding and staffing are covered. We have a strong evidence base and can demonstrate these things. However, although we're all integrated and harmonised, the rest of the University hasn't caught up with us yet."
Stage 5	**EXTERNALISE / EMBED**	Very high confidence. Everyone in our Department / Research Group is doing this, and it is embedded in our workflow to the point we don't have to think about it. All new staff who join the group comply with this. No staff members are left out or overlooked. We are harmonised with the rest of the Institution. We may be working, where appropriate, to agreed external standards. We are working, where appropriate, with others outside the University.

2.2.4 Crowdsourcing

In order to try to capture ideas from the user community, we have deployed an experimental crowdsourcing technique on the project website – www.southamptondata.org. Crowdsourcing is essentially an online *suggestions box* through which visitors can also vote with either a positive (thumbs up) or negative (thumbs down) response for each proposed idea, with the website maintaining a cumulative score for each. Here we are using a free trial of the Idea Scale online service (http://ideascale.com/). Figure 1 shows how this is embedded on to the home page, and how the results can be browsed and voted on.

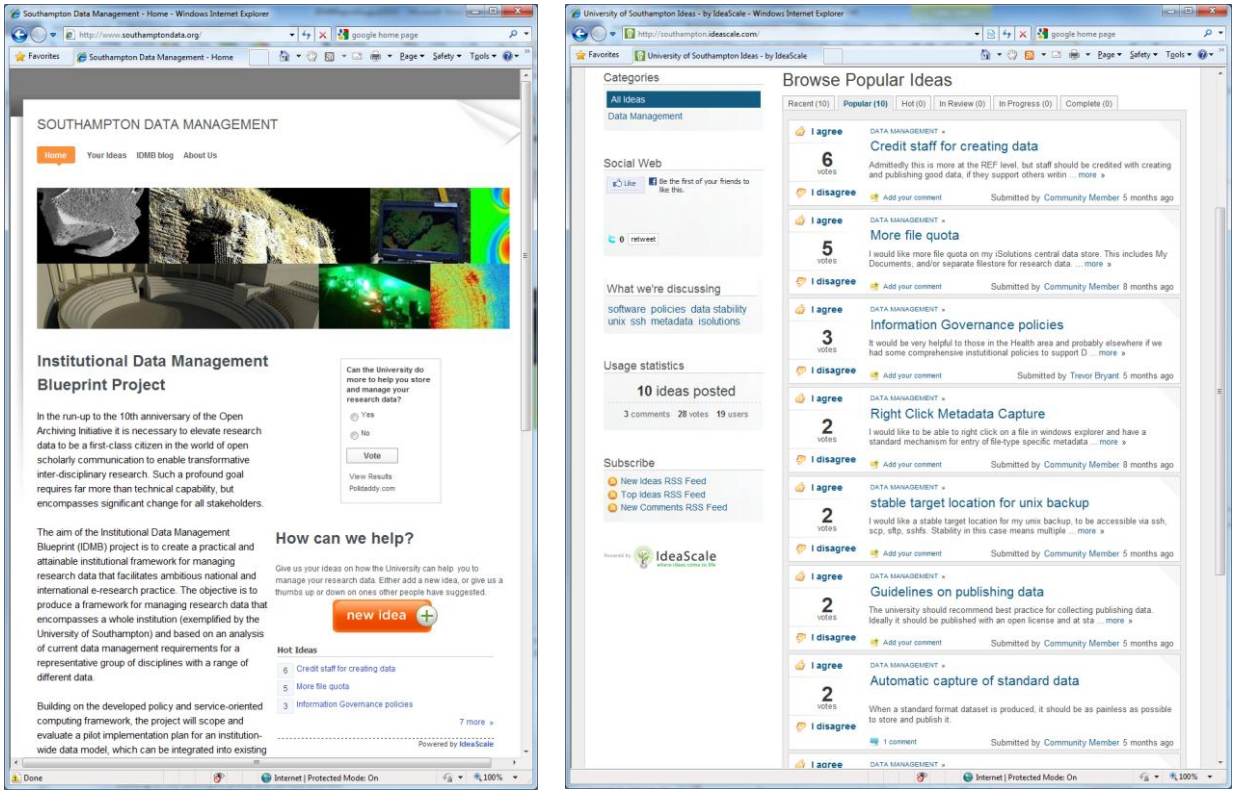

Figure 29. Crowdsourcing from project home page (left), and results (right)

While only a few ideas have been posted, they do seem to be of merit, and have been voted on. Crowdsourcing results, as of 7 September 2010, are shown in Table 12. We aim to promote this platform more in future to try and elicit more response from the community. As it is a low/zero cost task, we feel that it is worth piloting to see if it is worth further investment in the future for this, and other, projects.

Table 12. Crowdsourcing Ideas

Idea	Votes
Credit staff for creating data. Admittedly this is more at the REF level, but staff should be credited with creating and publishing good data, if they support others writing good papers.	6
More file quota. I would like more file quota on my iSolutions central data store. This includes My Documents, and/or separate filestore for research data.	2
Information Governance policies. It would be very helpful to those in the Health area and probably elsewhere if we had some comprehensive institutional policies to support Data Management infrastructure.	3
Right Click Metadata Capture. I would like to be able to right click on a file in windows explorer and have a standard mechanism for entry of file-type specific metadata that is stored in the file if possible but also in an external network metadata store.	2
Stable target location for UNIX backup. I would like a stable target location for my unix backup, to be accessible via ssh, scp, sftp, sshfs. Stability in this case means multiple years/decades long term. I would like to point to the backup target once when I arrive and then forget about the backup.	2
Automatic capture of standard data. When a standard format dataset is produced, it should be as painless as possible to store and publish it.	2
Single well managed research data repository. Rather than have lots of little sites, keep all the research data in one big tool. Make it easy to get the data out so we can build subject specific overlays and tools, but manage curation centrally.	2
Encourage use of machine readable datasets. PDF and other "rendered" formats are not very useful for reuse.Excel, CSV & RDF are much more useful for the long term.	0
Archive Software as well as data. If bespoke software was used to produce data or results (graphs etc) then the software should be archived to allow the work to be repeated. If it requires anything beyond a known baseline linux/windows then perhaps the entire stack (OS/libs/app) should be stored. If the software is compiled, then the method of compilation should be preserved. The same also goes for software required to interpret or view some data then that should be preserved too. There's the oft told story of the NASA data carefully preserved, but nobody thought to keep a copy of the data structure or software used to view it! JISC have an interest in this area; http://www.software.ac.uk/	0

2.3 Summary

In this section we have described the data management audit carried out within the Schools of Archaeology, Electronics & Computer Science, and Engineering Sciences. An online questionnaire was supplemented with face-to-face interviews in order to drill-down on particular topics to obtain a fuller picture. In summary, the following are some of the key points obtained:

- Guidance and advice on research data management was limited;
- Knowledge of available capability and resources was limited;
- Researchers resorted to their own best efforts in many cases, e.g. USB hard drives;
- Data requirements are growing, almost half of respondents stored more than 100GB of research data;
- Most users had experienced problems due to lack of storage;
- Longevity of storage is considerable, mean 5 years, many researchers express preference for keeping research data 'forever';
- Backup practices were inconsistent, with users wanting better support for this;
- Researchers need help on how to organise their research data;
- Many researchers share, or would like to share, their data;
- Many researchers use other people's data, particularly within their own group;
- There is considerable scope for improvement in the provision of resources and capability.

A modified version of the AIDA toolkit was used to perform benchmarking of the current status of research data management in the three Schools surveyed. While some concerns over the validity of the process for completing the AIDA survey were expressed, it has proven useful as a basic check.

In addition to the findings above from the questionnaires and interviews, the following could be inferred from the AIDA process:

- Capabilities across different Schools varies, with pockets of best practice throughout;
- Schools research practice is embedded and unified;
- Most of the data management capability tends to be localized;
- Formalization of data management policies and procedures would be beneficial;
- Technological capability needs to be more uniformly supported at the institutional level;
- Resources are generally limited.

These preliminary findings will be augmented with results from other Schools and Faculties across the University of Southampton, for the final report.

3 Kick-off Workshop

A workshop was held at the University of Southampton on 24 March 2010 to engage the researcher community in the IDMB project. We had over 60 delegates from 10 of 23 academic schools attend. We also had key representatives from the Research & Innovation Services, iSolutions, Library and IT Innovation (university spin-out company).

The project has high-level backing from the University of Southampton's senior management, and the workshop was opened by Professor Phil Nelson, Deputy Vice-Chancellor for Research. The audience was briefed on the aims of the IDMB project, and there was general agreement that this topic was important to the University and the attendee's roles.

3.1 Outputs

The aim of the workshop was to elicit feedback from the attendees on their current views on data management support at the University, and to provide some insight to the project on what directions would be beneficial for researchers, and related staff, at the University.

After an overview of the project was presented, a general discussion was opened out. Key issues raised in this discussion are shown in Table 13.

Table 13. Kick-off workshop discussion

Question:	Discussion:
To what extent do Research Councils (RCs) understand the cost implications of data management?	The University is responsible for its own data. Some RCs require you to upload data into an archive. Some RCs give part of your grant for data management. Soon, all RCs will have a data management policy.
What legal issues are being investigated in IDMB?	This comes under the governance IDMB work package and is being explored. Conditions may be put on data sets. S3RI[12] are working on data confidentiality. Need to think about how you will manage the data before you start your project.
Is software data?	Yes, although issues regarding ownership and IPR must be covered carefully.

[12] Southampton Statistical Sciences Research Institute (S3RI) - http://www.southampton.ac.uk/s3ri/

A more structured brainstorming session followed during which attendees were asked to highlight:

- **Quick wins.** Actions that can help in the short-term with data management.
- **Dreams & aspirations.** Long-term goals that could make a deep impact on the productivity and visibility of research data and processes.
- **Issues & frustrations**. Problems and barriers to being able to practice good data management.

The collated answers to these three questions are shown in Table 14.

Table 14. Kick-off workshop outputs

Quick Wins	Dreams & Aspirations	Issues & frustrations
Ask Tesco/BP how they do it! University standards / protocol guidelines. Data management pro-forma. Seminars about data management. Advice on data labelling. Make Graduate School data management course compulsory. Are people aware of existing data services at the university? No quotas for file storage. Secure data more quickly and risk analysis of data loss. Stop single point failure of hard drives. Easy access to repositories across departments. Repository for data management plans.	Visibility of my research history and related colleagues / students. Seamless integration from papers to source data. Central data body linking all university / RC archives together. Intuitive and natural system, making putting data in and getting it out "child's play". Automated archive and meta-data generation. Knowledge base for sharing research data and resources – including bibliography lists, external data sets, contacts etc. Expert support for planning and implementing data management plans Kudos for good data management. Integrated policy approach by RCs internationally. Capturing data in electronic form, without overhead (e-lab note books).	Metadata: how to make it meaningful and useful later is hard to define upfront. Ongoing curation costs and implications are not fully understood. Responsibility of data generated by former university members. Researcher's exclusive use of own data: guidelines, times etc. Viewing / reprocessing data if you do not have the original program installed. Multiple institution / university guidelines. Legal issues; FOI requests indicate that the public "owns" the data, not the university. Security: Who has access and where? What are the implications of a compromise? Who is responsible? Should we curate bad data? Lost data.

3.2 Summary

The kick-off workshop was well attended by a wide cross-section of the University population, highlighting the importance with which this subject is held institutionally. The level of engagement was high during the workshop and provides additional data that is consistent with the results of the data management audit described in Section 3.

4 Data Management Framework

4.1 Policy, Governance and Legal Issues

Research data management is a core capability of any research organisation, such as the University of Southampton. The management of research data is fundamental to good practice, and it is expected that all those engaged in research follow appropriate procedures. With current and future legislation and drivers to open up data, it is important to consider what is required to support researchers. In this section we describe the current governance framework that exists at the University, including both internal and external drivers. Compliance is discussed and recommendations are provided in order to improve the communication and implementation of data governance at the University of Southampton to better support its researchers.

4.1.1 Internal Governance

This section describes the current internal governance structures surrounding research data management at the University of Southampton.

4.1.1.1 Research Integrity and Academic Conduct

Researchers at the University of Southampton are treated professionally and a code of conduct is in place that sets out the standards by which they are expected to adhere to. This document is openly published [1] and covers the following areas:

- Leadership and organisation
- Academic conduct
- Academic Fraud
- Documenting results and storing primary data
- Publication and its responsibilities
- Academic discourse
- Ethical conduct of research
- Refereeing
- Complaints

The practice documented here is generally pervasive across the institution, through Graduate School training and supervision to personal development of researchers. Of notable relevance to this study is the policy on *Documenting results and storing primary data*[13], which states:

"Throughout their work, it is good practice for researchers to keep full, clear, and secure records, whether in paper or electronic form, of their procedures and results, including interim findings where applicable. They should include accurate and contemporaneous records of primary experimental data and results, in a form that will provide clear and unambiguous answers to questions concerning the validity of data later. This is necessary both to demonstrate good research practice and to answer subsequent questions. Such **records should be kept for 10 years after collection or subsequent publication**, whichever is later." (report author's highlighting)

It is this area that the IDMB project is focussing on, in terms of how the institution can better support this.

[13] http://www.soton.ac.uk/ris/policies/integrity.html

4.1.1.2 Intellectual property

Intellectual property (IP) is knowledge created and includes inventions, literary and artistic works, and symbols, names, images, and designs used in commerce; IP is owned by the original creator. A summary of intellectual property rights is given in Table 15.

Table 15. Intellectual Property Rights

Registered Rights	**Patents.** Patents are a 20 year monopoly awarded by the State for patentable inventions. For an invention to be patentable it must be novel, inventive and capable of industrial application (and not in excluded categories). Patents can be obtained for a new product, a new process or method and in some circumstances new uses for an existing product. To obtain a patent an application must be filed and relevant fees paid to the government. **Registered Design Rights.** Registered Design Rights protect the appearance of the whole or part of a product, particularly its colour, shape, textures, lines and contours. The design must be new (not the same as one already in the public domain) and have individual character (must give a different overall impression to an informed person). Registered Design Rights last for 25 years (renewable every 5 years). **Registered Trade Marks.** Trade Marks are applied for through the Patent Office and are used to protect a word, logo or other symbol (including noises, smells and sounds) applied to or associated with classes of products or services. A registered Trade Mark is granted for an unlimited duration (providing it is renewed every 10 years) and is denoted by ®.
Unregistered Rights	**Copyright .** Applies to literary and dramatic work and is an automatic right, there is no need to apply. Works should be marked with the authors name, date created and ©. The right continues for 70 years after the death of the author. **Unregistered Design Rights.** Design right allows you to stop the copying of your design therefore you must be able to prove that it is your design that has been copied. There are no registration formalities, but you should record the design in a design document and be able to prove when the design was first created. Protection lasts for 15 years from this time. **Common Law Trade Marks.** Common Law Trade Marks are trade marks most commonly established through 'use' and/or having an established reputation. It is then likely that the mark is already used to sell or advertise goods or services for a period of time sufficient that the public comes to associate the Mark with those commodities. It must be proven that a rival using your mark intends to mislead or confuse consumers. **Know how, Trade secrets, Confidential information.** Value of the above Intellectual Property cannot be undervalued. It is the knowledge you have that adds something extra to a process or method and cannot be deduced from the end product. In most circumstances such information should only be revealed under confidentiality or non-disclosure agreements.

The University has obligations, under the terms of most grant funding, to both disseminate and foster exploitation of research results. Patenting is a valuable tool for promoting the commercial development of inventions that need sustainable financial investment. The University may also have obligations to its commercial and academic collaborators to co-operate in the protection and exploitation of research results. Obtaining patent protection, however may incur substantial external costs (£100,000s), with significant expenditure in the first three years. The University has a dedicated department (Research and Innovation Services) that can advise on the protection of IP and the data related to the specific IP.

4.1.2 External Drivers

In this section we describe the many external drivers that determine the University's behaviour with respect to data management and publication. This is a fluid area, especially with the advent of the Freedom of Information Act 2000, and the recent Climategate[14] issues. This section is tackles the following themed areas: funders, Climategate, data security, and Freedom of Information.

4.1.2.1 Research Councils

A significant proportion of the University's research is funded by Research Council's UK (RCUK), and similar overseas funding agencies. RCUK comprises:

- Arts and Humanities Research Council (AHRC)
- Biotechnology and Biological Sciences Research Council (BBSRC)
- Engineering and Physical Sciences Research Council (EPSRC)
- Economic and Social Research Council (ESRC)
- Medical Research Council (MRC)
- Natural Environment Research Council (NERC)
- Science and Technology Facilities Council (STFC)

These are non-departmental public bodies, accountable to Parliament, and are charged with funding science and research through investment of taxpayer's money.

Any research funded by RCUK must adhere to the guidelines set out by the individual Council providing the funding. Research Governance is the responsibility of the Research Organisation, including maintaining high standards of research integrity and methodology. This is described for the University of Southampton in Section 4.1.1.1 of this report.

In terms of data management, the different research councils have different detailed policies, but adhere to a general set of principles. In 2009 the Digital Curation Centre (DCC) produced a report entitled "A report on the range of policies required and related to digital curation" [3]. This report considered:

1. Curation requirements of UK research funders
2. Gaps in current curation policies
3. Recommendations of policy development

DCC quotes from a Research Information Network (RIN) report [4] that sets out the "five principles required for effective stewardship of digital research data":

1. The roles and responsibilities of researchers, research institutions and funders should be defined as clearly as possible, and they should collaboratively establish a framework of codes of practice to ensure that creators and users of research data are aware of and fulfil their responsibilities in accordance with these principles.

2. Digital research data should be created and collected in accordance with applicable international standards, and the processes for selecting those to be made available to others should include proper quality assurance.

3. Digital research data should be easy to find, and access should be provided in an environment which maximises ease of use; provides credit for and protects the rights of those who have gather or created data; and protects the rights of those who have legitimate interests in how data are made accessible and used.

4. The models and mechanisms for managing and providing access to digital research data must be both efficient and cost-effective in the use of public and other funds.

5. Digital research data of long-term value arising from current and future research should be preserved and remain accessible for current and future generations.

DCC researched the main UK research funders' policies and related support infrastructure; a summary is shown in Appendix IV. It can be seen that there is variation between the different funders, with EPSRC and STFC having few requirements relating to data curation. It is also interesting to note that guidance and monitoring is not well covered.

DCC believe that essentially, the funding bodies require two stipulations relating to research outputs:

1. Research outputs are created in an appropriate manner to ensure that they can be made widely accessible

2. They are maintained in the long-term to facilitate future access, either under the auspices of the institution in which the funded researcher is based or by means of deposit in a special repository or data centre.

DCC believes that significant action is required in developing institutional policies. They recommend that a group developing curation policies should consider:

1. Resources such as the OpenDOAR (http://www.opendoar.org/index.html) policy tool, UK Data Archive (http://www.data-archive.ac.uk/) and DCC guidance should be used where possible to help close gaps in the digital curation landscape. Making policies available will also help others build on best practice.

2. Policies need to be mindful of context: a data management plan for example needs to complement and work in harmony with the relevant institutional and repository requirements for curation.

3. Existing structures could be used to embed curation in research workflows, for example researchers could be directed to advice on data management as part of funding application procedures in the same way ethical approval is currently ensured.

4. Existing staff such as librarians, Freedom of Information officers or departmental representatives could take on a broader support role to act as curation champions and broker relations between staff and the various support services.

5. Attention should be paid to encouraging uptake of any new policy. Preliminary scoping exercises, test phases or a reward system that recognises researchers who adopt best practices could be useful.

6. Policies need to be practical and accompanied by the required support infrastructure to ensure they can be implemented. A mechanism to monitor implementation and revise the policy to amend inappropriate clauses is crucial.

The DCC has produced a "Checklist for a data management plan" [5]. They believe this will aid researchers when producing data management plans within proposals to funding bodies and the subsequent development should the grant application be successful.

DCC sum up by stating that there are three broad issues that will need to be addressed by all stakeholders "if we are to create a stable base from which to develop meaningful curation policies".

1. Identify roles and responsibilities for curation.

2. Assess cost and benefits to determine how and by whom curation should be financed.

3. Develop a robust and sustainable curation infrastructure with appropriately skilled staff.

4.1.2.2 European Commission

Andrew Smith from UKRO[15] was contacted concerning European Commission (EC) policies concerning data management. The EU does not have a specific policy regarding data management however in the current Seventh Research Framework Programme (FP7) the EC is running a pilot scheme initiative on open access to peer reviewed research articles Grant agreements in these areas signed after the beginning of the open access pilot will contain a special clause requiring beneficiaries [6]:

> 1. to deposit articles resulting from FP7 projects into an institutional or subject-based repository;
>
> 2. to make their best efforts to ensure open access to these articles within six months (Energy, Environment, Health, Information and Communication Technologies, Research Infrastructures) or twelve months (Science in Society, Socio-economic Sciences and Humanities).

There appears to be no specific requirement around data relating to research articles.

4.1.2.3 Technology Strategy Board (TSB)

Dr John Morlidge, who is a Lead Technologist at the TSB, was contacted to discuss any data management requirements of research grants. Dr Morlidge stated that there are no formal policy documents in place. The TSB would expect data management to be arranged between the project partners in the Collaboration Agreement and any IPR developed during the project.

The only mention of data is in the typical Project Offer Letter that states "the participant is responsible for producing all information, maintaining proper records, complying with the terms of any legislation or regulatory requirements and the TSB's terms and conditions of grant".

[15] http://www.ukro.ac.uk/

4.1.2.4 Industry

Research and Innovation Services at the University of Southampton handles industrial interactions. A discussion was held with Kevin Forshaw in R&IS concerning specific requests from industry concerning data management. Mr Forshaw has been involved in many contract negotiations with industry but could not recall a specific request regarding data management. However, the School of Engineering Sciences at UoS hosts a University Technology Centre for Computational Engineering funded by Rolls Royce. They have stipulated that all files relating to the centre are hosted on a separate, secure server. Other technology centres within SES and the University more widely, have similar agreements in place.

Within UoS, company specific data is typically covered by a confidentiality agreement (CDA). Within the CDA, UoS agrees to keep all the data relating to the project and not disclose for a period of 5 years but periods up to 20 years have been agreed.

4.1.3 Climategate

The Climate Research Unit at the University of East Anglia has recently been embroiled in controversy relating to the authenticity and access to research data (dubbed "Climategate"). Thousands of e-mails and data were leaked following the hacking of a University server and articles soon appeared in the media suggesting data irregularities [7,8]. When asked for information under the Freedom of Information Act (FOI), the scientists could not produce location data from Chinese weather stations and it appeared that e-mails suggested manipulation of the data. Furthermore, it was alleged that CRU had been obstructive to data requests under the FOI Act. Investigations by the House of Commons' Science and Technology Select Committee [9] and an independent Science Assessment Panel commissioned by the UEA [10] concluded that there was no evidence of malpractice on the part of the CRU and Professor Phil Jones though they did find that there was room for improvement in some of the CRU's working practices.

This case highlights issues surrounding how data is managed and subsequently shared as well as the need for guidelines on institutional data management.

4.1.4 Data Security

The School of Medicine at UoS is particularly concerned with information security where records of patients are involved. It is also possible that personal data may be held by Schools such as Health Sciences, Law and Humanities. For example, all applications to the Integrated Research Application System (IRAS) must provide details and copies of policies in the National Information Governance Board (NIGB) section of the application. These include:

- Compliance with information security standards
- Details of Data Protection Registration
- A copy of the Information Security Policy
- A copy of the Network Security Policy
- A reference copy of any System Level Security Policies

Similar information regarding information security may also be requested by Pharmaceutical companies who are sponsoring research activities within the School of Medicine.

Additional information may be required regarding the implementation of these policies and how the following are addressed:

- Security and Audit Measures
- Physical Security

- Network security
- System Security Risk Review
- System Monitoring
- Encryption
- Data Retention & Destruction

Currently within UoS there is no coherent set of policies that a third party can be directed to.

Table 16 shows a mix of Regulations, Guidelines and Policy documents that cover information security at UoS but some do not appear to have been reviewed for over three years.

Table 16. Information Security policies and guidance

Documents around Information Security	URL	Date last reviewed
Main Index to documents – University Regulations	http://www.southampton.ac.uk/isolutions/regs/university/index.html	01-Sep-09
Regulations for the use of Computers and Voice and Data Communications Networks	http://www.calendar.soton.ac.uk/sectionIV/computers.html	28-Aug-09
Guidelines for Management and Use of Software and Licensed Data Products	http://www.southampton.ac.uk/isolutions/regs/university/softdata.html	18-Feb-09
Guidelines for Access to and Dissemination of Information through the Internet	http://www.southampton.ac.uk/isolutions/regs/university/inet.html	24-Aug-06
Model regulations for facilities use	http://www.southampton.ac.uk/isolutions/regs/university/model.html	24-Aug-06
University data network — terms and conditions	http://www.soton.ac.uk/inf/termsandconditions.shtml	01-Sep-09
Regulations for Use of iSolutions Services Resources	http://www.calendar.soton.ac.uk/sectionIV/isolutions-resources.html	30-Sep-09
Public workstation area use policy	http://www.southampton.ac.uk/isolutions/regs/isolutions/workstationarea.html	15-Feb-07
Legal & ethical use of facilities	http://www.southampton.ac.uk/isolutions/regs/isolutions/ethical.html	30-Sep-08
Data Protection Policy	http://www.soton.ac.uk/inf/dppolicy.pdf	30-Sep-08
Electronic Communications Policy	http://www.southampton.ac.uk/isolutions/regs/university/ECommsPolicy.html	

Other institutions are more advanced than UoS in developing and publishing their security information. Examples of good practice from other Universities are listed in Table 17.

Some of these institutions have used the Universities and Colleges Information Systems Association (UCISA) toolkit Information Security Edition 3.0 [11]. The toolkit presents the components required to assemble policies to meet local needs and meet the ISO20071:2005 standard.

Table 17. University security documentation

University of Birmingham	http://www.it.bham.ac.uk/policy/documents/UoBInformationSecurityPolicy.pdf
University of Leeds	http://campus.leeds.ac.uk/isms/information_security/index.htm
University of Leicester	http://www2.le.ac.uk/offices/itservices/resources/cis/iso/Policy-Documents
University of Reading	http://www.reading.ac.uk/internal/its/help/its-help-pcsecurity/its-Information-Security.aspx

4.1.5 Freedom of Information Act

The Freedom of Information Act (FOIA) came into enforcement in 2005. Freedom of Information Ltd [12] state that "the FOIA has had a profound affect on the public and private sectors alike, and provides companies, campaigners, journalists and citizens with a powerful new tool. Since 2005 everyone has, for the first time, a legal right of access to information". It should be noted that certain data is exempt such as if disclosure compromises national security or is commercially sensitive.

An article by the BBC [13] investigated a request by Douglas Keenan under the FOIA to Queen's University in Belfast to release over 40 years of research data concerning tree rings. The University suggested a number of reasons for not disclosing the information such as commercial confidentiality and intellectual property rights. However, a legal ruling by the Information Commissioner concluded that Queen's had wrongly used legal exemptions to withhold the data requested and could be in contempt of court if they did not hand over the data. At the time of publication the BBC stated that "Queen's is now considering its position". Douglas Keenan is a sceptic of scientists who claim global warming is a result of human activity. Below is an excerpt from his website [14]:

"Some people have asked why Queen's University Belfast (QUB) does not want to disclose the data. In fact, most tree-ring laboratories do not make their data available: it is not just QUB and Gothenburg that have been reluctant. The reason for this was elucidated by Peter M. Brown, in April 2007. At the time, Brown was president of the Tree-Ring Society, which is the main international organization for tree-ring researchers. Following is an excerpt.

… they ARE my data. Funding agencies pay me for my expertise, my imagination, and my insights to be able to make some advance in our understanding of how nature works, not for raw data sets. … It is the understanding and inferences supplied by the scientist that funding agencies are interested in, not her or his raw data.

In other words, even if the research and the researcher's salary are fully paid for by the public—as is the case at QUB—the researcher still regards the data as his or her personal property. Baillie confirmed this in an interview with Times Higher Education *in July 2010, saying "As far as we were concerned, it was our data … the data belonged to the people who made the measurements".*

Much of the controversy surrounding Climategate was concerned with refusals to disclose information under the FOIA. A report by the House of Commons Science and Technology Committee [9] states "we can sympathise with Professor Jones, who must have found it frustrating to handle requests for data that he knew, or perceived, were motivated by a desire to simply undermine his work". They also add that his "blunt refusal" to disclose scientific data "was in line with common practice in the climate science community". However, the report goes on to conclude "we cannot

reach a firm conclusion on the basis of the evidence we took but we must put on record our concern about the manner in which UEA allowed the CRU to handle FOIA requests. UEA needs to review its policy towards the FOIA and reassess how it can support academics whose expertise in this area is limited".

The University of Southampton does have guidance on the FOIA and a form where third parties can request information [15].

4.2 Services and Infrastructure

4.2.1 Current Services and Infrastructure

This section summarises the as-is infrastructure provided at the University of Southampton, and in particular from the central ICT organisation (iSolutions) and the Schools of Chemistry, Engineering Sciences and Electronics and Computer Science, in support of research data management.

4.2.2 Data Storage and Management Facilities

The University of Southampton has recently consolidated its IT staff centrally, from a more devolved structure. The central ICT department (iSolutions) offers a number of facilities for the storage and some degree of management of various forms of data, shown in

Table 18. This table is comprehensive, and includes systems targeted at education, but which are also sometimes used for research data.

Each School also has its own facilities, and these are described in more detail in Section 4.2.2.1.

Table 18. Central (iSolutions) data storage and management facilities[16]

Facility	Capacity/ store	Capacity /total	Accessibility	Structure of Data	Notes
Personal Filestore "My Documents"	10s GB	10s TB	Individual	Low	No formal quota, but expectation that volumes will be low. Designed for individual, personal and non-collaborative data
"Resource" Filestore	100s GB	10s TB	School	Low	Fixed size, designed for administrative documentation shared within a school of group.
"Research" Filestore	Low TB	100s TB	Defined group	Low	Intended for storage of large data sets specific to individual research projects. Shared amongst an identified set of individuals.
IRIDIS Scratch Space	10s TB	10s TB	Individual	Low	Very short term temporary storage for working data sets under computational processing. Not backed up.
Mediabin	10s GB	Low TB	Defined group	Medium	Web interface, primarily for

[16] Note that "Capacity (per store)" refers to the size of an individual data set, rather than overall capacity of the facility.

					web content management.
Content Management System	Low GB	100s GB	Global	Medium	Content is globally accessible when published
Document Management System	10s GB	100s GB	Defined group	High	Any file format is supported, but the DMS is primarily a repository for reports, spreadsheets and similar office-style data. Aimed towards administrative activity rather than research data. Collaborative amongst a defined set of individuals.
Subversion	10s GB – 100GB	100s GB	Defined group	Medium	Primarily aimed at source code or other version-controlled text, but can be and is used for storage of other data types.
EPrints	Low GB	100s GB	Global	High	Primarily documents and papers, though multimedia supported
Portal Groups	100s MB	100s GB	Defined group	Low	Shared files via SUSSED portal
SharePoint	10s GB	100s GB	Global	High	Extranet, and pilots for research data
Local PC ("C:\local\")	100s GB	100s GB	Individual	Low	Not backed up, local storage only.
Oracle RDBMS	10s GB	Low TBs	Any	High	Data access is controlled by custom applications written by the end user.
Microsoft SQL Server	10s GB	Low TBs	Any	High	Data access is controlled by custom applications written by the end user.
LAMP MySQL	10s – 100GB	100s GB	Any	High	Data access is controlled by custom web applications written by the end user.
Blackboard	Low GBs	Low TBs	Defined group	Medium	Educational course content
Perceptions	100s MB	100s GB	Individual	Low	Primarily exam scripts and associated media
Moodle	100s MB	100s GB	Defined group	Low	Educational course content

4.2.2.1 Schools ICT

In this section the IT infrastructure related to research data for the School of Electronics & Computer Science, Engineering Sciences, and Humanities (Archaeology) are described. This is in the context of the centrally provided capabilities described in section 4.2.2, and the recent transitioning of staff and resources from Schools to management by iSolutions.

4.2.2.1.1 School of Electronics and Computer Science

The School of Electronics and Computer Science's nine research groups carry out a broad spectrum of research from core computer science, through to electrical and nano-engineering. Requirements across these areas vary and the School has a number of systems to support their researchers (Table 19). As computing is core to their research there is significant School capability and expertise. Researchers use the School facilities in the main, rather than central ones, and these are supported by a team of dedicated staff.

4.2.2.1.2 School of Engineering Sciences

The School of Engineering Sciences has historically had its own IT systems, but these are gradually being transitioned to iSolutions. The School system comprises three main file servers: two of which are relatively old and of limited (100GB) capacity. The main School file server currently has a capacity of 10TB and is used for research, administrative and teaching support.

Researchers also have their own servers used for a variety of different roles. There is a significant amount of computational research in the School, including eScience projects, that require the use and development of specialist systems. Several researchers also make use of national HPC facilities, including leading the UK Turbulence Consortium. Laboratory work also requires considerable data storage in many instances, and local solutions are often used. Examples of researcher's own systems include:

- Computational results from HPC, for research community use (publicly accessible);
- Research file store for industry-funded technology centres;
- Experimental results for research community use;
- Computational results and source code for researcher team;

Each of these cases involves several terabytes of storage requirement, in some cases several tens of terabytes. It is estimated that at least 50TB of data storage is locally provided by researchers and their teams in SES, although this is a conservative estimate.

Researchers often take it upon themselves to purchase their own storage, either as local (e.g. USB) hard drives, or servers. This cost has historically been born by the researchers in many cases, but this has meant that market rates are leveraged, which is cost-effective at the time of purchase but may have downstream consequences.

Researchers distinguish between reproducible and non-reproducible data. In the former case, setup files and logs are often stored centrally (and hence backed up), so that data can be reproduced. For the latter case users may be more willing to pay a premium for secure, backed up services.

The issue of data publication arises here, as several research teams host data for their communities. This has many benefits for the University, in terms of raising research profile, but is not supported fully at present. It is an area that promises significant return on investment in terms of research assessment, and visibility of the University's research.

Table 19. School of Electronics & Computer Science data storage and management facilities[17]

Facility	Capacity/ store	Capacity /total	Accessibility	Structure of Data	Notes
Personal Filestore "My Documents"	Staff 3Tb Students 16Tb	20 TB	Individual	Low	Quotas enforced, but requests for more quota are routinely agreed. Designed for individual, personal and non-collaborative data
"Resource" Filestore	100s GB	10s TB	School	Low	Fixed size, designed for administrative documentation shared within a school of group.
"Research" Filestore	1 TB per project	100s TB	Defined group	Low	Bought by projects. Intended for storage of large data sets specific to individual research projects. Shared amongst an identified set of individuals.
Content Management System	Low GB	100s GB	Global	Medium	Content is globally accessible when published. School Bespoke CMS & wiki database tables. Includes *docpot*, a local 'media bin' and and simple Document Management system.
Subversion/ ECS "Forge"	10s GB – 100GB	100s GB	Defined group	Medium	Primarily aimed at source code.
EPrints	10s GB	100s GB	Global	High	Primarily documents and papers, though other formats supported

[17] Note that "Capacity (per store)" refers to the size of an individual data set, rather than overall capacity of the facility.

4.2.2.1.3 School of Humanities (Archaeology)

In this section we describe the Archaeology researchers within the School of Humanities. These users rely heavily on centrally provided IT infrastructure, while also having specialist requirements to satisfy.

Extensive use is made of the central iSolutions *My Documents* and *Resource file stores* by all researchers. The Archaeological Computing group makes particular use of HPC facilities, and hence uses the attached *IRIDIS* scratch space for temporary storage of large working files.

A number of projects use the University's media asset management system, MediaBin, for photos and videos, although the experience is not ideal. This is also used to share media for marketing purposes. EPrints is used for research outputs, as for all Schools, but is not currently used for management or publication of data.

Specialist systems are used for specific research projects, including:

- Autodesk Vault, for CAD and 3D modelling data;
- Integrated Archaeological Database (IADB);
- Portfolio, used for photographic data management;
- Web servers, LAMP, for project sites including research data, and online journal;
- Other systems used, e.g. NOC and ECS.

It is notable that the Archaeology researchers were unaware of certain central facilities (e.g. "Research" Filestore,) and in other cases have found central facilities insufficient to meet their full range of needs, e.g. need for a Portfolio system rather than MediaBin.

4.2.3 Data Infrastructure Analysis

In order to visualise the current IT infrastructure, Figure 30 shows each data store plotted against its ability to share data (x-axis) and the degree of structure of the data (y-axis). Here we include iSolutions and Archaeology systems, to illustrate the typical overlap between central and School systems. The size of the bubble represents the maximum size of an individual data set within the data store, and the arrows represent the total capacity of the data store, to the same logarithmic scale. The colour represents the degree of difficulty to the end user in storing their data in this store; systems that require the production of large amounts of structured meta-data, for example, will be more red than green.

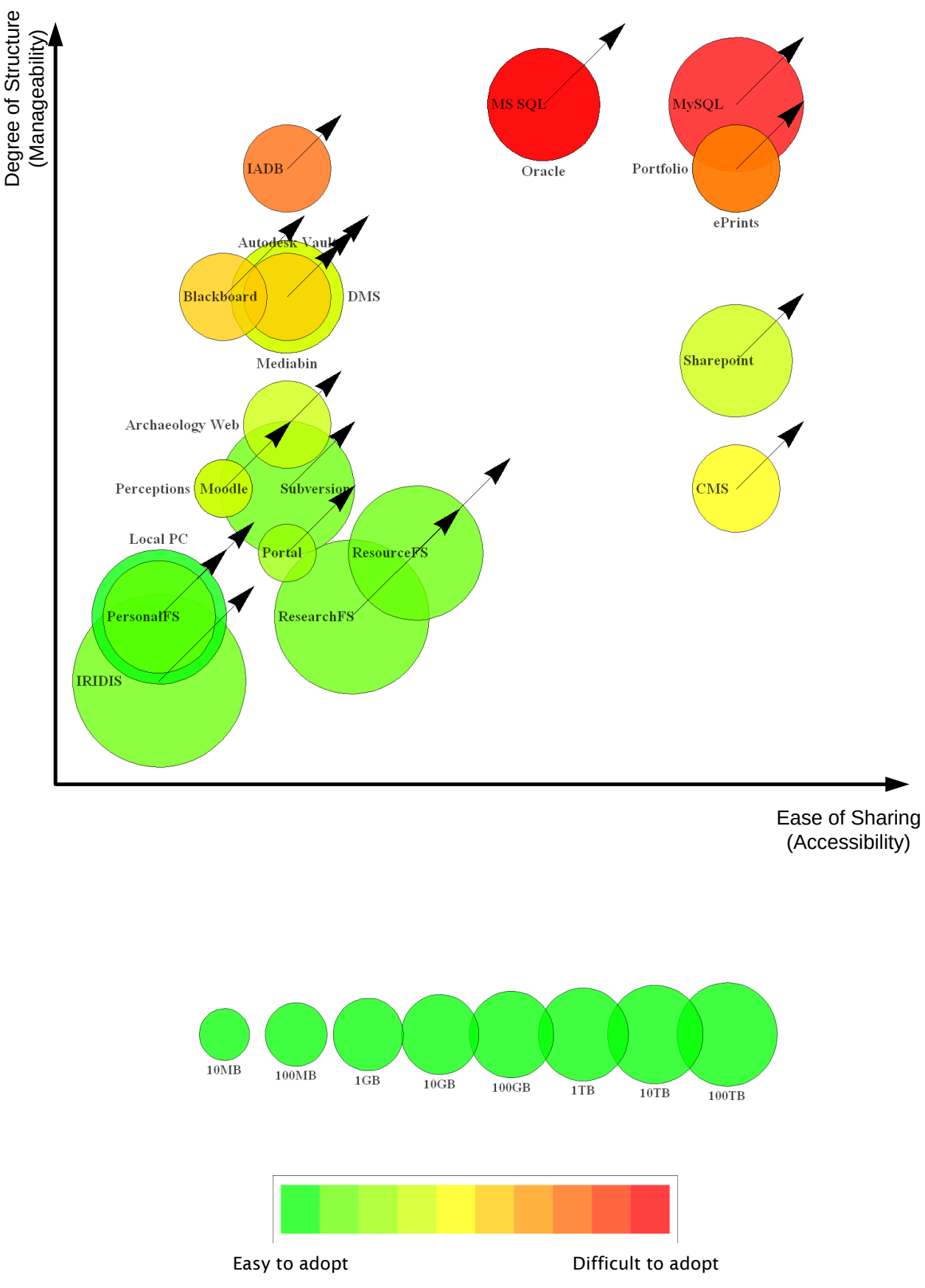

Figure 30. Current data management systems

A number of conclusions can be drawn from this analysis.

1. Large data stores tend to be unstructured

In general, any data store provided by the University which is high volume is also essentially in the form of a flat file system. The University does offer some database systems which can provide a high degree of structure, but these are both more difficult to adopt, and whilst they offer good capacity on a per-user basis, they may not be scaled for large numbers of users.

2. Large data stores tend not to be shared

The larger data stores provided by the University are normally in the form of individual or small group file stores, which do not lend themselves to sharing and collaboration with a wider group. Again the exceptions to this are the database systems, which can be made globally accessible through the development of web interfaces to the data; however this clearly suffers the caveat of difficulty of adoption. Further, as already noted, these systems are currently only scaled for a small number of users. User access control tends to be controlled centrally.

3. Structured data stores are often difficult to adopt

This is perhaps self-evident – the higher the degree of structure in the data store, in general, the more effort is required to position that data within the structure, or to produce and record appropriate meta-data.

The converse is also true – unstructured data stores are easy to adopt – one simply stores the data in any way, anywhere.

4. Structured data stores tend to be task-specific

This is not evident from the diagram above, but by examination of the tables. In general, data stores which have been designed to be well-structured rely upon format-specific meta-data or characteristics in order to organise the data. For example, "MediaBin" understands well how to organise images, just as "ePrints" understands well how to organise papers and documents. Again the exceptions are the pure database systems, which require proprietary coding and therefore have a high barrier to adoption. Structured data stores can be considered to be more like end-user applications.

5. There is no coherent approach to data storage and management

There are a large number of data stores available throughout the University. This appears to be a consequence of needing to provide a greater degree of structure to data, and accessibility of data. As a result, medium-specific data stores have proliferated in addition to the more general, but less structured and accessible flat file system data stores. As such, there is no single point at which to take a view of the data owned by the University, and therefore no coherent methodology to finding and accessing data.

4.2.4 Discussion

Data storage and management offerings have evolved at the University from an initial provision of simple, low capacity flat file system storage for individuals and University schools. As demand has grown, so to has the capacity of these data stores, consequentially the majority of data storage capacity in the University is in the form of flat file systems. This is provided both centrally and locally by researchers who do so as the most cost-effective means of satisfying their particular needs.

As the need to organise, index, and share data has arisen, the University has implemented a number of point solutions to meet these needs for different data types. These data stores tend to be of relatively low capacity, and since these solutions generally rely on an understanding of the underlying format of the data being stored, they can be restrictive in the type of data that can be usefully stored. As a result, there is a proliferation of these systems.

There is no single view of the data held by the organisation, and therefore no simple mechanism by which to search or make accessible this data. There is no coherent approach to how or where data should be stored, and no understanding of the quantity and range of data which is held by the University. Clearly guidance is needed in this area.

In order for the university to be able to create a sustainable data management infrastructure, capacity must be increased so that it can satisfy demand. This must be coupled to a business model that is both manageable by the University and attractive to the research community, both in terms of usability and cost.

It is interesting to note that this is the same situation that evolved around High Performance Computing at the University over the last decade. The University has successfully evolved from having a devolved, disorganised HPC capability, to a consolidated and sustainable capability that is world-leading.

4.3 Gap Analysis

In this section we discuss the area in which the University can help its researchers to improve their data management capabilities. It is based on the data management audit described in Sections 2, 3 and 5.

4.3.1 Policy, guidance and legal

The University does set out policies and guidance related to data management, including the legal framework for issues such as intellectual property rights, confidentiality and freedom of information. This is handled largely by the Research and Innovation Services team.

Availability and accessibility of policy and guidance is lacking

It is apparent, however, that researchers are either not aware, or unable to easily access this information. The information is scattered in different places, and researchers unable to find it. Guidance is not always clear, and not always in the right format.

Legal, policy and budgetary information and responsibility is not clear to researchers

In all of the data audit activities it is apparent that researchers are not clear as to the legal, policy and budgetary responsibilities to them. While best practice is shared, and in general research integrity is upheld in the highest regard, how this relates to data management is not always clear. Notably, it appears that this where this information is available, it is not in a format that is easily understandable by all researchers. The overall relationship between researchers and the institution in terms of rights and ownership is not understood by all.

Advice to researchers is not coherent or readily available

In terms of practice, researchers seem reliant on their local support networks for advice. While this is valuable, this means that there is no coherent or consistent advice on data management. In particular, little central information or advice seems to be made available to the individual researcher at their desktop.

Points of contact for data management guidance are not clear

Where researchers cannot find data management solutions locally, they find it difficult to know who to turn to. The central IT service, iSolutions, has a single support channel (ServiceLine) but it appears that researchers are reluctant to use this for their higher-level data management queries.

Available Support is not always apparent to researchers

There is a a disconnect between the support made available by the institution, and what researchers think is available. It appears that a significant number of researchers are not aware that the data management problems they are suffering can easily be solved by the institution. The most notable example is that researchers believe that there is a limit to central disk storage, while iSolutions provides effectively unlimited support (within reason). This demonstrates a lack of effective communication between support services and researchers.

Areas of responsibility between groups is not always clear: Schools, Faculty, iSolutions, Library

There is a lack of coherence and clarity of responsibility between local (School, Faculty) and central (iSolutions, library) staff and facilities. The consolidation of IT at the University of Southampton is clarifying some of this, but it is still rather unclear as to the responsibilities at the boundaries between academic and professional services when it comes to data management.

Lack of guidance on open and shared data, and freedom of information

The current push towards open and shared data, along with implications of Freedom of Information in the wake of the Climategate situation requires some analysis so that appropriate guidance for researchers can be given.

4.3.2 Infrastructure and services

Data management is up to the end user

Ultimately the researcher must design their own data management solution. This is, however, difficult when there is a general lack of central guidance and support; as described in Section 4.3.1. This means that researchers must find out themselves using their local community of practice, which can lead to more local solutions that cause problems.

Lack of coherent data management structure that users can understand (lack of communication)

It is clear that there is no clear data management strategy or structure that is coherently communicated to researchers. This causes confusion and can lead to researchers taking it upon themselves to find a local solution, for example USB external hard drives. This creates risks and also is not always optimal in terms of researchers' time and efforts.

Proliferation of systems means that researchers can find it difficult to navigate options

Due to the organic growth of data management solutions at the University, there is a proliferation of different systems, from local and shared file systems, to managed repositories. This plethora of systems can be confusing for researchers, and also means that infrastructure investments may be more diluted than is optimal. There do not appear to be clear data repository solutions that the majority of researchers can easily take advantage of.

Lack of capacity

Researchers believe that there is a lack of capacity for server data storage. This leads to them finding local solutions, typically USB external hard drives or desktop NAS systems. More storage is available from iSolutions than many researchers realise, notably *ResearchFS* shared file system. However, if researchers were to apply for the amount of storage they currently hold locally, then this may be overwhelmed. As one researcher suggested, giving every PhD student a 1-2TB external hard drive would solve their local problem, but scaling this to university-wide server solution would be costly. Clearly, the latter does need serious consideration, as local storage solutions are not secure or sustainable.

Lack of scalable business model for infrastructure and services

It is apparent that the business model for providing data management solutions to researchers at an affordable level is not scalable at an institutional level. This has led to local solutions being procured by researchers, and an inability for iSolutions to deliver in a manageable way. Without a business model that is attractive to both researchers, and the University, it is difficult/impossible to improve data management capability and realise the benefits.

Automated backup for researchers' data

While the University does provide automatic backup services, many researchers do not take advantage of these because they either do not know about them, find them hard to use, or there is insufficient capacity for their data. This situation needs to improve, with better communication a good starting point.

Data lifecycle not clearly handled – particularly archive and curation of researchers' data

It is unclear as to how research data can be archived for curation and preservation. This issue is prevalent when, for example, an academic retires and the data that they have is of benefit to the research community, both locally and internationally.

Lack of data management tools

The availability and knowledge of data management tools is limited, particularly with regard to the issues described above. This makes data management a largely manual and ad-hoc process, which is not ideal.

4.3.3 Training and practice

Little formal training around data management

Data management as a separate topic tends to be embedded into research methodology but not covered explicitly. This means that best practice is not disseminated, often leading to local, ad-hoc solutions.

Lack of support for Data Management Plans

There appears to be little support for the creation of data management plans for research projects. As this requirement becomes more prevalent, researchers will need more support.

Little self-help and guidance

Researchers have found it difficult to find help and guidance when they need it on the subject of data management. Their main recourse is colleagues and ServiceLine requests to iSolutions. There is little central high-level guidance on best practices for data management.

Practice varies greatly across disciplines, and between researchers

There are few standard patterns and practices that can be shared between researchers, and across disciplines, that the university supports centrally. Best practice tends to be passed down locally. This leads to the data management ecosystem developing in an ad-hoc manner, making it difficult to improve in a systematic way.

Little consideration for sharing and preserving data

While there are some excellent examples of data sharing and preservation, such as eCrystals and the Open Data work led by ECS, there is no general framework for enabling this. The issue of preservation and curation for research data is also one that needs addressing as data becomes more important in support of research outputs in the future.

4.4 Metadata Strategy

Metadata is generally defined as *data about data* and has been used for centuries to organise information. In the digital domain this becomes more important for discovering information. In the context of research data management, metadata has a number of different uses. In general, we can define three categories of metadata [32]:

- Descriptive – title, author, keywords
- Structural – how objects are connected (e.g. datasets connected to articles)
- Administrative – Who can access, when/how created, file type
    - Rights Management (IPR)
    - Preservation Metadata

Mostly we are concerned with Descriptive Metadata which includes free-text fields (title, abstract, etc.) with various standards in use (e.g. Dublin Core, various standards used in libraries AACR2/RDF/MARC21). Increasingly, connectivity is becoming important as the semantic web and linked data concepts are becoming reality – including extensive research around ontologies.

In addition there are various forms of keywords, that can include Thesauri, Classifications, Authority Lists [28] – e.g. Library of Congress Subject Headings, Art and Architecture Thesaurus (AAT). Thesauri, word lists & subject headings will give a controlled list of terms often arranged hierarchically. Sometimes thesauri are used to generate subject headings by automatically including any broader terms (e.g. buildings – houses – cottages). This can improve retrieval, especially if the search system doesn't automatically include broader/related terms/etc.

Classifications (e.g. Dewey Decimal, Library of Congress) are similar to Subject Headings and traditionally have been used to order items on a shelf, as opposed to for information retrieval. While many subject headings would be applied to an item, only one classification would be assigned, since it could only be in one place. In a digital environment this makes less sense and classifications tend to be more rigid. However since they are numbers (or alpha-numeric strings) they can be more machine readable.

Authority Lists (e.g. Library of Congress Authorities, Union List of Artist Names (ULAN)) are usually lists of people, places and organisations, sometimes part of Thesauri/Subject Headings but often separate for practical reasons.

Also there can be unstructured keywords – often referred to as a folksonomy (e.g. tagging used in Flickr, Delicious, etc.) – that are often used to add meaning.

4.4.1 Metadata user scenarios

In order to understand how people use metadata, a significant amount of research has been carried out around digital photo storage. We investigate this briefly in order to gain insight into human nature and how metadata is used for large collections of shared resources.

A common example where people voluntarily add metadata is storing photos on the web site 'Flickr'[18], both titles/captions and tags. Angus [17] found 86% of images had tags, with a mode average of four (looking at images in groups with 'university' in the name). 39% of the tags were 'selfish' – only useful to the individual and their families/friends, while 52% would be useful for the wider community using Flickr. Common tags include year, location, season, lighting, features (sky, sea) and colours [19].

[18] www.flickr.com

It is interesting to note the motivations for adding tags [16]:

- Improved retrieval
- Sharing with friends and family
- Forming ad-hoc photo pools – i.e. by using the same tag it can be easier to find photos uploaded by others about the event etc. A similar purpose to use of hashtags in twitter.

Professional photographers / serious amateurs use the site differently – being less concerned about using tags to find images. Their key motivation is to generate interest / feedback on their work, browsing latest images from their contacts, and photostreams of people who comment on their work [18,19].

Tagging as a manual process incurs an overhead on the user, so the question of automatic metadata generation is an interesting one. In many cases with text-based items automatic metadata generation may be more effective. Heckner [20] found journal articles tagged in Connotea (online tool to organise and share references) often used words also in the full-text of article (54% identical, 16% variations). Keywords chosen by the authors tended to be more specific and higher in number.

However automatic tagging of images based data is difficult [35] – when no text based title/caption is available. Proposed work on new CAPTCHA systems to distinguish humans from web-bots assumes automatic tagging of images is impossible. There are systems aiming to do this (e.g. http://alipr.com), however they often assign incorrect terms. Hollink [21] refers to a semantic gap between high level concepts users search for an low-level features (colour, shape texture) searchable by Content based retrieval systems There are evidence systems that suggest likely terms, can increase the number of tags people use (e.g. ZoneTag application for Camera Phones [16]). There is research on automatic image tagging but few genuine automatic tagging products around. For example, alipr and semi automated tools like LabelMe from MIT.

Arts and humanities disciplines make extensive use of imagery and of temporal and spatial systems of classification. Image and spatial metadata have an extensive, well developed system of metadata structure and attribution. Temporal metadata remains poorly developed but an emerging field of interest. The following sections explore the metadata strategies appropriate in these domains.

There is significant advice on managing images from JISC [26], key to this is understanding the user's requirements. Shatford [38] identified three categories of description, building on Panofsky's earlier work [33]: Generic (e.g. women); Specific (e.g. Mona Lisa) and Abstract (e.g. beauty). Each of these can be divided into Who / What / Where / When. In a later work [37] Biographical (history/life of the image), Exemplified (type of image it is an example of) and Relationship (how image relates to other versions/formats).

JISC recommends using existing metadata standards or guidelines where they exist within the community, and to adapt and existing standard where they don't [27]. Advantages of using standards include resource discovery and sustainability.

JISC describe a variety of options for systems to manage images from personal systems (e.g. Adobe Bridge), web-based services (e.g. Flickr), commercial/open source/bespoke image management systems and collection management systems (e.g. Eprints) [25]. Which option to adopt will depend on our needs and constraints.

Advice on file formats (JISC 2006) suggests a non-proprietary open standard (e.g. TIFF) for archiving original images, and various formats for delivery depending on needs.

4.4.1.1 Metadata user scenarios for research

In terms of research data the scope for metadata is wide. In order to understand how different disciplines are approaching this, we consider four typical user scenarios where metadata can add value:

- **Bring me all the research data related to *X*.** Where *X* is a metadata tag, such as author, date, subject, or other attribute.
- **Bring me all the data that relate to a given research publication.** In this scenario the user has found a publication of interest and would like to drill down into the data. This could be done directly, or through a search using the known metadata associated with the publication.
- **Bring me all publications related to this piece of research data.** In this case the user has found an interesting piece of data, and they would like to see if there are any publications relating to it.
- **Bring me all of the data related to this piece of data.** In this case the user wants to find data related to a set of data that they have found.

In all these cases metadata would make the tasks easier. The exact format of this metadata, and additional semantic functionality, would enable better searching and findability.

With these user scenarios in mind, we now present a number of discipline-specific examples of how metadata is used by researchers.

4.4.1.2 Archaeology

In Archaeology there is not a single preferred format or metadata scheme. Key organisations provide standards and guidance and a variety of schemes and formats are used depending on the type of data. Examples of some of the key organisations that drive development and adoption of metadata standards are described here:

- CIDOC CRM (conceptual reference model) enables mapping between both metadata and data across archaeological records. Applications include:
- English Heritage STAR project - specifically for excavation data (see also ArchaeoTools and CultureSampo for related approaches)
- The Archaeology Data Service (ADS) is the UK's repository for archaeological data, funded by the Arts and Humanities Research Council. It includes extensive documentation on archaeological metadata capture and formats.
- Forum for Information Standards in Heritage (FISH)
- INSCRIPTION wordlists (including Thesaurus of Monument Types)
- MIDAS XML (commonly used for Monuments scale data)
- Online archaeological data management systems including IADB, ARK and INTRASIS. Some of these enable web service exposure of content and metadata (one example developed at University of Southampton)
- Getty Art and Architecture Thesaurus
- Collections Trust's Archaeological Objects Thesaurus
- UKOLN

In terms of metadata formats, these are dependent on the type of data being described.

- Temporal period metadata. This began with the English Heritage Timelines Thesaurus, with nascent projects including COMMONERAS and international collaborations with colleagues via the CAA-Semantic-SIG. The latter has a great deal of expertise in cultural heritage semantic web activity and was established by University of Southampton researchers.
- Image assets in archaeology. Commonly documented according to JISC DIGITAL MEDIA guidelines. DUBLIN CORE (in all forms) is seen habitually in archaeological image data records alongside extracted EXIF, XMP or IPTC data. Also Adobe Bridge/ MediaBin (proprietary format); AutoDesk Vault – CAD files.
- Geospatial data. Commonly attributed via the UK GEMINI (Geo-spatial Metadata Interoperability Initiative) metadata elements. Metadata management in UK GEMINI is supported by Geographic Information Systems habitually used with archaeological research. The ArcGIS ArcCatalog (proprietary format) is also used.
- Geophysical data. Metadata strategies for archaeology are supported by the English Heritage Geophysical Survey Database
- Laser scanning data for cultural heritage metadata. This is dealt with by Laser Scanning Addendum to the Metric Survey Specification[19].

There is a need for workflow management as part of the metadata for a range of archaeological data types. For example, polygonising, registering and cleaning laser scan data transform data in considerable ways. Similarly geophysical data undergo a range of transformations in one or more software packages. There is a need to provide a repeatable process.

Archaeology has no consistent toolsets for the attribution of metadata. Thus, for image data some users adjust cameras in order to predefine certain IPTC and EXIF data such as creator or GPS co-ordinates. Geophysical and other prospection devices may provide automated metadata but the bulk is ascribed in software during processing or must make use of external metadata created to accompany versions. No consistent tools are commonly made use of to achieve this. Individuals derive their own data management strategies and many do not record the necessary metadata to recreate the precise analytical outputs. Work with Geographic Information Systems is equally variable. ArcGIS and other systems enable the attribution of structured metadata but they do not generally document transformations in data in an automated way. Again such transformations could only be repeated with the creation of versions and versioning metadata.

[19] Heritage3d.org

4.4.1.3 EPSRC UK National Crystallography Service

Crystallography has well defined metadata requirements. This is described in more detail in the I2S2 project[20], with a brief overview being given here.

One example of a crystallography data repository is eCrystals[21] at the University of Southampton. This uses the eBank-UK Metadata Application Profile [29]. This Application Profile (AP) is encoded in the XML schema language (XSD). Broadly speaking, the profile records the following information:

Simple Dublin Core

- Crystal structure
- Title (Systematic IUPAC Name)
- Authors
- Affiliation
- Creation Date

Qualified Dublin Core (for additional chemical metadata)

- Empirical formula
- International Chemical Identifier (InChI)
- Compound Class and Keywords

The repository uses Digital Object Identifiers [34] as a form of reference identifier as well as the IUPAC International Chemical Identifier (InChi) IUPAC [23] as a domain identifier.

On DLS beam line I19, raw data is collected and ingested into the central data storage facilities at STFC using the General Data Acquisition (GDA) system which was developed in-house.

The Core Scientific Metadata Model (CSMD) [31] was developed to help organise data derived in investigations using the large-scale facilities at STFC. This model captures the experiment proposal and the acquisition of data, relating data objects to each other and their experimental parameters. The CSMD is currently used in the ICAT suite of tools [22]; a data management infrastructure developed by STFC for the DLS and ISIS facilities. ICAT is primarily intended as a mechanism to store and organise raw data and for facility users to have systematic access to their own data and keep a record for the long term.

Processed and derived data are normally taken off site on laptops or removable drives and the results data are independently worked up by individual scientists at their home institution. STFC makes no provision for data storage and management other than for raw data generated in-house.

4.4.1.4 Materials Science

In materials science metadata is being increasingly used to classify materials and testing in order to create repositories for both research and industry use. A survey of metadata needs has been carried out as part of the Materials Data Centre project [36]; the MatDB schema (represented in Figure 31, Figure 32 and Figure 33) was discussed with two researchers.

- The source section (who generated the data and a link to the report) was most useful, two additional fields were suggested (Supervisor and Research Group) and to use a template since much of the data is likely to be reused in multiple datasets.

[20] http://www.ukoln.ac.uk/projects/I2S2/

[21] http://www.ukoln.ac.uk/projects/I2S2/

- The usefulness of elements of the Materials section was considered debatable, and the ability to add a microstructure image of the material being tested was desired. A less rigid production element was desired, even a free-text box.
- Of the Test section, the Test Condition element was daunting, and it was suggested optional values should be hidden from the user.

The suggested method was a MatDB toolbox based on an online web form. Motivations for spending the time to creating a compliant file were identified as:

- The ability to store data that you can refer to in the future
- Allowing other people to access your data
- You wanting access to other people's data
- Being able to find data related to yours

While the MatDB schema is comprehensive, feedback from researchers was that it was complicated, and this was a barrier to widespread adoption within existing workflows. It demonstrates the overheads that detailed metadata assignment can incur.

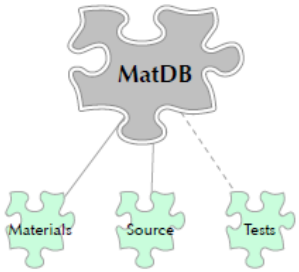

Figure 31. MatDB schema.

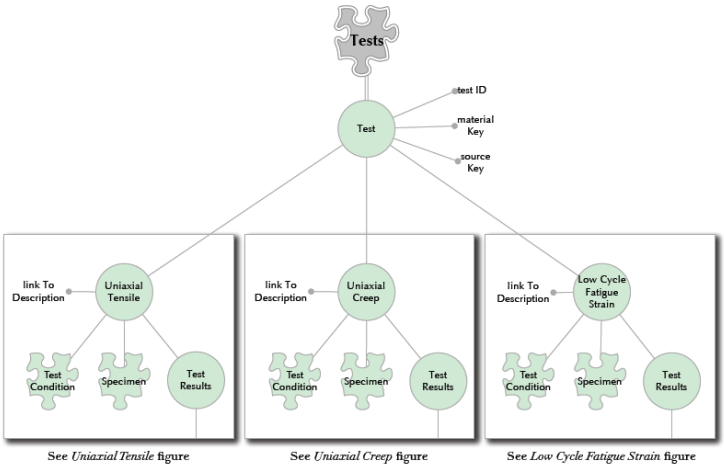

Figure 32. Materials test types for MatDB schema.

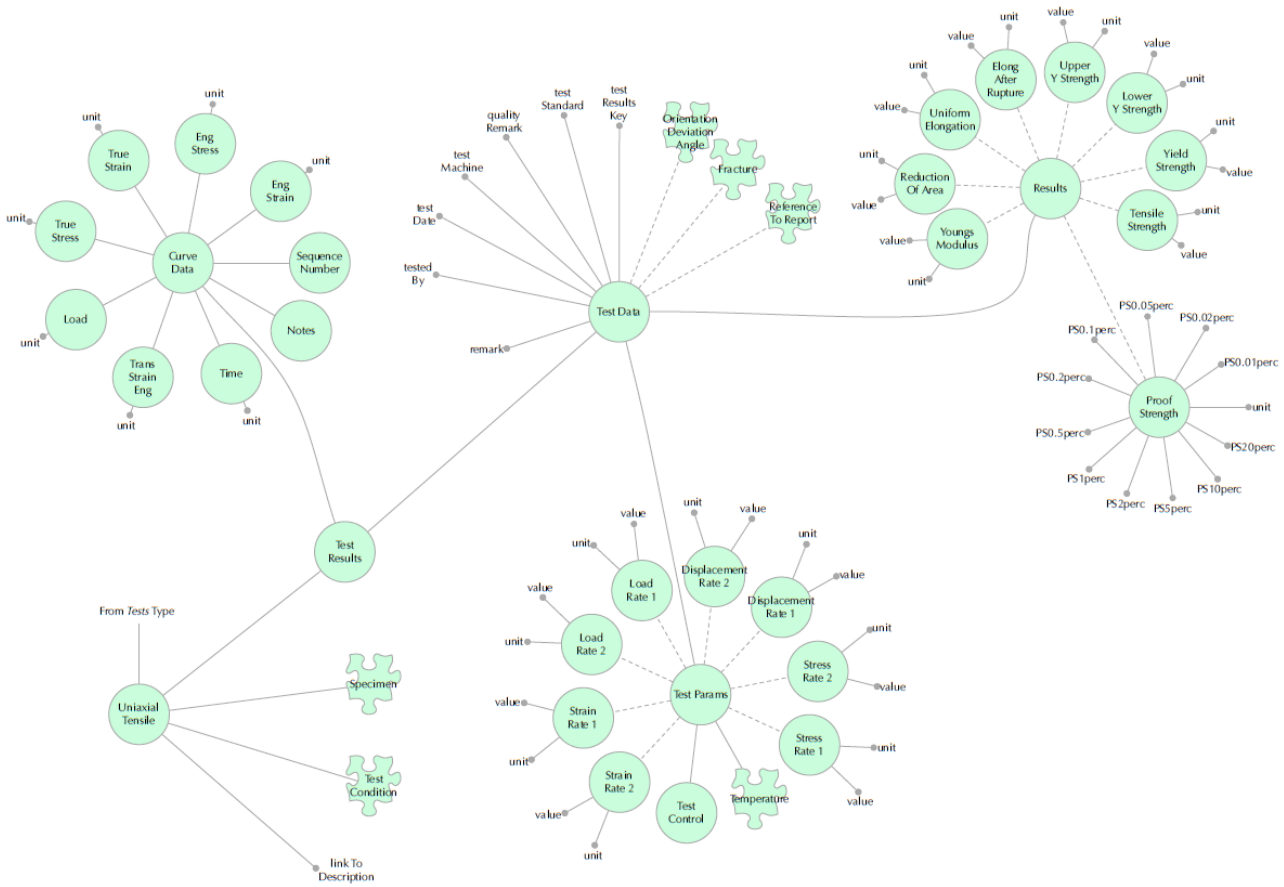

Figure 33. Uniaxial tensile test for MatDB schema.

4.4.2 Metadata framework

The aim of the IDMB project is to provide some guidance and pilots for better data management across the University of Southampton. As described above, the use of metadata is necessary to add meaning and context to research data. Our photo example described in 4.4.1 does highlight, however, that both input and usage of metadata varies depending on both the creator's intent, and end user's desires.

In order to try and create a metadata framework that is applicable across domains, we look at what the end user would want to achieve. As an example, we use the analogy of archiving a retiring professor's office in a day – there is not enough time to record/classify each object. Instead papers, etc. of a similar nature are placed into folders, and then into a box, with a label describing the contents in general terms. In future researchers can access material if they identify a box file. Once opened they can identify folders, and then, if of sufficient interest, they can then find the relevant papers to investigate in detail.

We break this down into three levels of findability:

- Core metadata *(box file)*. In order to find author, publisher, discipline, date;
- Discipline metadata *(folder)*. To find the right sub-domain, project, funder, technique;
- Project metadata *(paper)*. To find detailed dataset and its context.

This three layer metadata approach is illustrated in Figure 34 with discipline examples for aeronautics and archaeology shown in Figure 35. If used appropriately we believe that this metadata model is able to satisfy the requirements of the user scenarios identified in 4.4.1.1. This model provides flexibility for the creator, while trying to include applicability to the end user. The difficulty occurs when these people have different roles, such as a researcher as the creator, and an archivist as the end user.

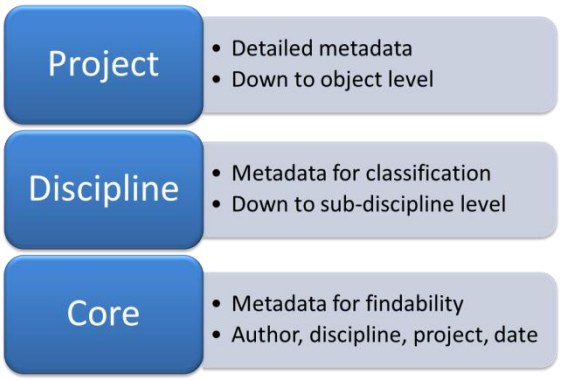

Figure 34. Three-layer metadata model

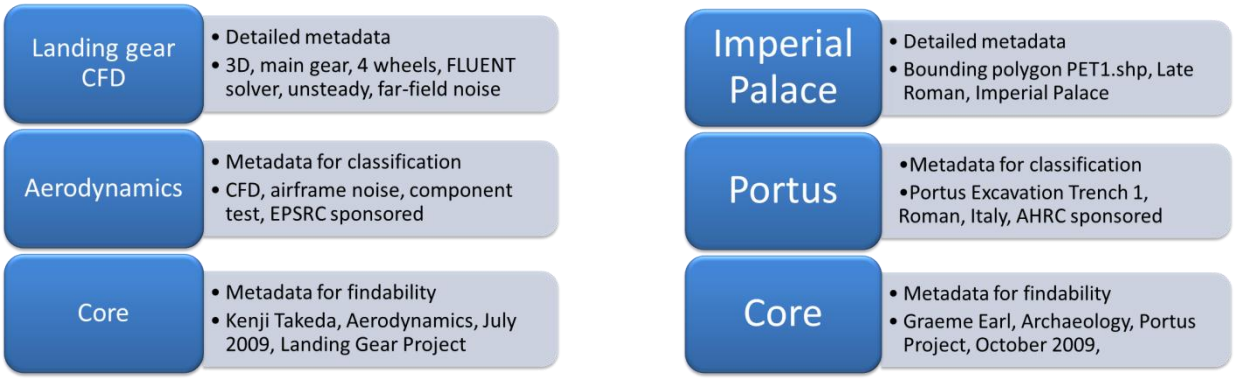

Figure 35. Three-layer metadata model examples for aeronautics (left) and archaeology (right).

One of the more developed standards for metadata is Dublin Core, which is a set of text-based elements that can be used to describe resources such as books, media, and data. It has been developed since 1995 and now defined as ISO standard 15836 and NISO Standard Z39.85-2007. The continued development of this is now through the Dublin Core Metadata Initiative (DCMI[22]).

Simple Dublin Core comprises of fifteen basic elements, described in Table 20. Qualified Dublin Core has three additional elements to cover: audience; provenance and RightsHolder. Dublin Core is typically implemented in XML, and as such has been popularised for open access through the Open Archives Initiative for Metadata Harvesting (OAI-PMH)[23].

Here we propose that Dublin Core is an appropriate standard to be used for an institution-wide metadata framework to provide the first-level of metadata. This is the approach already used by the National Crystallography Centre at Southampton through eCrystals (see 4.4.1.3), and is supported by EPrints.

[22] http://dublincore.org/

[23] http://www.openarchives.org/OAI/openarchivesprotocol.html

Key challenges that face the uptake of better metadata management include:

- How to encourage people to tag their data;
- Metadata schemas that are not onerous;
- Usable tools for metadata assignment and import;
- Provenance tracking;
- Automating metadata assignment.

These will be all developed and addressed as part of the pilot studies in archaeology and the nano-fabrication centre, described in Section 5.

Table 20. Dublin Core elements[24]

Attribute	Description
Title	The name given to the resource. Typically, a Title will be a name by which the resource is formally known.
Creator	An entity primarily responsible for making the content of the resource. Examples of a Creator include a person, an organization, or a service. Typically the name of the Creator should be used to indicate the entity.
Subject	The topic of the content of the resource. Typically, a Subject will be expressed as keywords or key phrases or classification codes that describe the topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.
Description	An account of the content of the resource. Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.
Publisher	The entity responsible for making the resource available. Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.
Contributor	An entity responsible for making contributions to the content of the resource. Examples of a Contributor include a person, an organization or a service. Typically, the name of a Contributor should be used to indicate the entity.
Date	A date associated with an event in the life cycle of the resource. Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [Date and Time Formats, W3C Note, http://www.w3.org/TR/NOTE- datetime] and follows the YYYY-MM-DD format.

[24] http://dublincore.org/documents/usageguide/elements.shtml

Type	The nature or genre of the content of the resource. Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the DCMIType vocabulary). To describe the physical or digital manifestation of the resource, use the FORMAT element.
Format	The physical or digital manifestation of the resource. Typically, Format may include the media-type or dimensions of the resource. Examples of dimensions include size and duration. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource.
Identifier	An unambiguous reference to the resource within a given context. Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Examples of formal identification systems include the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).
Source	A Reference to a resource from which the present resource is derived. The present resource may be derived from the Source resource in whole or part. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.
Language	A language of the intellectual content of the resource. Recommended best practice for the values of the Language element is defined by RFC 3066 [RFC 3066, http://www.ietf.org/rfc/ rfc3066.txt] which, in conjunction with ISO 639 [ISO 639, http://www.oasis- open.org/cover/iso639a.html]), defines two- and three-letter primary language tags with optional subtags. Examples include "en" or "eng" for English, "akk" for Akkadian, and "en-GB" for English used in the United Kingdom.
Relation	A reference to a related resource. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.
Coverage	The extent or scope of the content of the resource. Coverage will typically include spatial location (a place name or geographic co-ordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [Getty Thesaurus of Geographic Names, http://www. getty.edu/research/tools/vocabulary/tgn/]). Where appropriate, named places or time periods should be used in preference to numeric identifiers such as sets of co-ordinates or date ranges.
Rights	Information about rights held in and over the resource. Typically a Rights element will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the rights element is absent, no assumptions can be made about the status of these and other rights with respect to the resource.

4.5 Summary

In this section we have highlighted the current data management framework at the University of Southampton.

It is clear that there is a robust policy framework at the University of Southampton that encompasses data management, ownership, IPR and freedom of information. The current issue is that this information is scattered in a way that it is difficult for researchers to access in a coherent way. It therefore appears that guidance is disjointed, and that the policy framework is not coherent. Similarly, guidance on data management is not clearly signposted, points of contact not clearly identified, and areas of responsibility between professional services not readily apparent.

Data management infrastructure and services are being consolidated within iSolutions, although Schools still house local capability of significant capacity. There is a plethora of different data solutions, coupled with a general lack of capacity. This has stimulated researchers to find their own solutions. While some of these are being migrated to central systems, the cost of doing this across the board would be significant.

The provision of backup services that are affordable and easy-to-use is not readily apparent to researchers. Also, researchers make the distinction between reproducible and non-reproducible data, which they are willing to manage cost/risk against in terms of paying for services. This is not readily supported at an institutional level.

It is apparent that while central systems can provide better support, there will always be more specialised requirements. Therefore the future strategy must combine commonality for consistent and affordable solutions, with flexibility to meet researchers varying needs.

The data lifecycle, particularly for curation and preservation, is not clearly handled by the institution, both in technically or organisationally.

There is a lack of formal training around data management, and limited self-help and guidance for researchers. In some areas there is exemplary best practice, and it is important that this is shared and promoted across the university for the benefit of all. This is important as data management plans become more prevalent. This will help to ensure that researchers are as productive as possible, in order to meet the university's ambitious strategic goals.

In terms of metadata management, an institution-wide framework has been proposed around a three-layer metadata model. The use of Dublin Core, and development of qualified Dublin Core, is suggested as a way of standardising use of metadata while providing extensibility within disciplines. Using a common framework has advantages in terms of training and support across the institution, development and use of tools, and embedding common data management practice within the researcher's daily lives.

Most importantly, there appears to be no coherent data management approach, with the current business model not being scalable, nor sustainable, to meet the current and future demand required to support the university's strategic goals.

5 Pilot Implementations

The initial phases of the IDMB project have highlighted many of the data management challenges faced by the institution. We have identified ways forwards, and will be embarking on three pilot implementations to see how more coherent, integrated, and intuitive data management solutions can be developed and deployed. These are in the areas of archaeology, the Nanofabrication Centre, and meta-search across federated repositories.

5.1 Archaeology

Archaeology researchers handle many different types of data, and cover a wide spectrum of requirements for data management that are applicable in other disciplines. In this pilot we will be exploring the use of Microsoft SharePoint 2010 as a virtual research environment, supporting researchers' data management needs.

In order to identify requirements, we have developed several user scenarios. It is envisaged that these will be delivered via a single platform.

5.1.1 User Scenario 1: Working with Geophysical Survey Data

David has a good understanding of technology. He has been doing geophysical surveys for many years and has developed his own best practice in working with the project data produced. Some of his work is commercial and some research and therefore IP and other restrictions need to be carefully dealt with.

During his archaeological fieldwork he keeps all the survey files produced on his laptop. The software packages he is using require specific file structures on his C-drive. This has caused him some problems, as at the Archaeological Computing Lab, where there are more powerful computers to process data, he does not have sufficient user rights to keep files on the C-drive. David is very busy trying to finish his analysis and writing reports so there is not much time left for proper archiving after the project fieldwork is completed. He deposits all the project files on the iSolutions Archaeology shared project folder, which is backed up every night.

He works very closely with colleagues in Italy. As he is responsible for the project he would like to have access to their data and vice versa, but he has not found a good solution for that.

David is generally happy about the way of working with his project data, even though sometimes he dreams about an assistant who helps to keep his digital and paper archives better organised.

5.1.2 User Scenario 2: Working with Computer Graphics

Tim knows a lot about 3D graphics and building survey. After receiving his Master's degree he worked as a research assistant. During that time he was involved in several survey and 3D computer reconstruction projects. He manages his data by capturing surveys either on the instrument or direct to a laptop. These are then backed up to a memory stick and frequently emailed home whilst on fieldwork. Gathering the data in two different ways can lead to confusion. Also each CAD file produced frequently contains the previous days' surveys in addition to that produced on the current day. The 3D reconstruction work builds on the surveys and produced a wide range of architectural and landscape simulations. Whilst there are a range of proposed methods for documenting the processes and files involved there is no consensus on the best approach. Currently Tim uses a hand written survey notebook and a typed 'diary' during his reconstruction work as the main documentation.

5.1.3 User Scenario 3: PhD Student Working from Home

Alison is a PhD student who mainly works at home. She finished her MSc of Archaeological Computing course several years ago. She also did her undergraduate degree in Southampton, so she is familiar with the infrastructure that the University provides.

She has her data very well organised: she keeps paper copies of readings and notes, and also backs up all her files regularly on her external hard-drive. When she comes to work in university she brings her laptop with her, or sometimes she copies files to her G-Mail account as an attachment. Alison does not use iSolutions disk space as she finds it difficult and unreliable to access from home. Also, a few times when she tried this method all of her GIS data files somehow became corrupted.

Alison makes use of online link and bibliographic management tools such as Zotero, Mendeley and Delicious. However, she feels that many of her colleagues would benefit from these but do not make use of them.

5.1.4 User Scenario 4: Senior Lecturer

Katie is senior lecturer. She works on an AHRC funded research field project. The AHRC requires deposit of a project archive with the Archaeology Data Service. As a consequence the data produced during the project must be consistent with deposit requirements, or be sufficiently documented and organised to enable the production of appropriate data and metadata with minimal additional investment. Preparation of data for deposit is rarely if ever built in as a major work package into Katie's work. As Katie wants to continue to develop her data beyond the lifecycle of her funded project she wants a means to expose her ongoing work in a way that makes it accessible and useful to others. As a consequence she requires that her data must as far as possible be exposed as RDF. Katie has three main needs for the documentation and management of her archaeological project data:

Image metadata:

On Katie's large field project image metadata is stored in EXIF and IPTC data and in an external metadata catalogue which enables CSV export. EXIF and IPTC data is attributed either direct via the camera (e.g. GPS spatial data, timestamp, creator, camera number) or via software such as AdobeBridge or download/ upload processing scripts. An automated tagging process or series of processes are run and checked prior to manual tagging. These data are managed centrally at the University of Southampton via a SharePoint server, using the Sharepoint Media Asset Library. At the end of each season data are uploaded to SharePoint. At the beginning of each season the data are downloaded from SharePoint and the data are locked on the server.

Temporal metadata:

Temporal data are gathered for many items on Katie's AHRC field project. A classic example is provided by Amphora data. Amphora sherds are recorded in the site office in Italy and entered into a database. The database will record the type of amphora. This type is in turn associated with various kinds of temporal information. The data and assets associated with these data should be attributed with appropriate temporal metadata and we will need to perform Allen operator based probabilistic reasoning ideally.

Spatial metadata:

Spatial data are also common on the project. Geographic Information Systems data include spatial information. The SharePoint server must ingest these data and enable their display. SharePoint is used to manage appropriate hierarchical spatial metadata. Note: the ArcGIS plugin can be used to

attribute data with a location using a map as with Flickr geocoding. Need to be able to see map data in any SharePoint page.

5.1.5 User Scenario 5: Retired Professor

Kris is an emeritus professor at the department. He retired a few years ago but he still filled with energy and ideas. Most of his life he has been using pen and paper and occasionally a typewriter, so he has a very big archive in his office. He admires technology but he remains a novice.

He use computer to write articles and to prepare his presentations. Kris has never lost any of his files even though he keeps them all in the My Documents folder with no further folder structure. He is very organised with his paper records, with them all nicely filed in his drawers and on his bookshelves. He can find a note on a paper record from twenty years ago but would find it nearly impossible to find a similar digital note.

In addition to Kris's paper notes he believes that he has the only copies of a number of vital paper documents. For example, the archives from a number of excavations remain in his filing cabinets. Similarly he knows of some physical archives from his and colleagues' excavations that are in stores in the department. Some of these are not organised and are poorly labelled.

One thing Kris desperately needs is an easy way of sharing the articles and presentations he has with other people. He has heard of ePrints but not used it yet. He also wants to make some simple web pages to accompany his existing publications so that he does not have to conventionally publish many hundreds of plans and photographs. He doesn't have any research budget to pay for this so ideally needs a system that he can learn to use himself.

5.2 Southampton Nanofabrication Centre

The second pilot will work with the Nanofabrication Centre (www.southampton-nanofab.com), the newly established state-of-the-art facility for nanofabrication and characterisation, run by the Nano Research Group from Electronics and Computer Science. An experimental data management repository will be established based on WP1 and WP2 for procedures related to two new pieces of equipment: the ASM Epsilon Epitaxy System and the Orion helium ion microscope[25]. Currently, although data from each experiment is stored digitally by the machines, records of the experimental settings used and the outputs obtained from both these systems are maintained in student logbooks; exploration of the parameter space is achieved by coordinated sharing of paper records. Initial discussions have shown that an eCrystals-style repository storing raw data, sufficient metadata to recreate the experimental conditions, plus data from the intermediate stages of analysis will have the capability to yield a significantly positive effect on the laboratory procedure.

5.2.1 User Scenario 1: Helium ion microscope single inspection

John is a researcher a new silicon device in the Southampton Nanofabrication Centre. All users of the facility use the computer-based Clean Room Management System (CRMS) to plan their experiments. They can choose from a *recipe book* and modify parameters to fit their task. Alternatively they can start a new process from scratch, entering all of the relevant machine and process parameters starting fresh. John chooses a standard process to being manufacture of his device.

John creates his device and then must inspect it using the Orion helium ion microscope. This again uses the CRMS, but here he revises the entries to what the process actually ran –i.e. actual parameters, rather than requested ones. The microscope produces a series of images, which are then transferred to an EPrints data system. EPrints requests the metadata from the CRMS so that it is stored alongside with microscope data. Once the experiment is finished, John can return to his office and access his data files over the network via EPrints. John can now manage his microscope data in an organised way without having to worry about storage, backup or archive, as this is now taken care of centrally.

5.3 Meta-Search

The third pilot will develop a proof-of-concept demonstrator that enables cross-disciplinary data linking and researcher expertise matching. This is to enable transformative inter-disciplinary science that is currently difficult to achieve due to the discipline-specific silo nature of open data repositories. The user community targeted is the Southampton Nano-Forum, a University Strategic Research Group, aligned to the national EPSRC theme, that comprises researchers across electronics and computer science, chemistry, physics, engineering sciences, and mathematics. It is based on the twin observations that (a) related disciplines tend to have similar working practices and academic values and (b) it is quicker to establish e-research and repository services at a departmental or subject level than an entire institution. This demonstrator will use the discipline-cluster common data model developed in WP1 and WP2 and apply a meta-schema across eCrystals, Materials Data Centre and the new nanofabrication centre repositories. This will provide an orthogonal view, compared to the conventional discipline-specific ones. As well as providing a mechanism for linking data, it will also be enabling social networking via data.

[25] http://www.ecs.soton.ac.uk/about/news/2607

6 Conclusions

The Institutional Data Management Blueprint project has carried out a data management audit across the School of Chemistry, Electronics and Computer Science, Engineering Sciences and Humanities using both top-down and bottom-up approaches. The conclusions from this provide insight into the current state of data management at the University of Southampton. This audit will be extended further in the next phase of the project. Notable conclusions so far include:

- There is a need from researchers to share data, both locally and globally;
- Data management is carried out on an ad-hoc basis in many cases;
- Researchers demand for storage is significant, and outstripping supply;
- Researchers resort to their own best efforts in many cases, to overcome lack of central support;
- Backup practices are not consistent, with users wanting better support for this;
- Researchers want to keep their data for a long time;
- Data curation and preservation is poorly supported;
- Schools research practice is embedded and unified;
- Schools data management capabilities vary widely.

In terms of gap analysis, the following major conclusions can be inferred:

- Policy and governance is robust, but is not communicated to researchers in the most accessible way;
- Services and infrastructure are in place, but lack capacity and coherence;
- There is a lack of training and guidance on data management.

It is apparent that there is no coherent data management approach, with the current business model not being scalable, nor sustainable, to meet the current and future demand required to support the university's strategic goals to deliver research excellence.

A three-layer metadata strategy based on Dublin Core has been proposed to provide a unified approach to improving data management across all disciplines.

Three pilot implementations around archaeology, the nano-fabrication centre, and meta-search across federated repositories, have been described and development work is starting on these.

It is clear that the current data management situation at the University of Southampton is analogous to the HPC landscape at Southampton a decade ago. The institution successfully moved to a more coordinated HPC framework since then that provides world-leading capability to researchers through a sustainable business model. A similar step change in data management capability is required in order to support researchers to achieve the University's ambitious strategic aims in the coming decade.

6.1 Recommendations

The data management audit and gap analysis indicates where improvements can be made in the short, medium and long-term to improve data management practices and capabilities at the University. The following preliminary recommendations are put forward for short (one year), medium (one to three years), long (more than three years) term action. The exact timing of implementation of recommendations is subject to further prioritisation by the institution.

6.1.1 Short-term

Crucial to supporting researchers is the consolidation of data management into a coherent framework that is easy to understand, use, and has a sustainable business model behind it. A number of major recommendations are put forward here for the short-term:

Create an institutional data repository

An institutional data repository would provide a coherent framework that could potentially satisfy the majority of researchers' data management requirements. This should be made usable so that it can be embedded into researchers' everyday practice. It should not preclude the development and deployment of more specialised capabilities, but should be extensible so that such systems could potentially be delivered through the central repository with customisation. It must have sufficient capacity and be affordable/free to attract users, rather than forcing them to develop/procure local solutions. This should be piloted, and then grown over the short-to-medium term.

Develop a scalable business model

In order to support data management in both the short- and long-term, a more scalable and sustainable business model needs to be developed. We have experience of this through development of HPC at Southampton, and there is work being done at other institutions to achieve this, e.g. Bristol, Edinburgh, and UCL. Without a clear business model and long-term commitment, it will be difficult/impossible to meet demand in an organised way.

One-stop shop for data management advice and guidance

There is a pressing need to provide researchers with the right information on policy, legal issues and guidance, so that they can rapidly create data management plans, and find out what is available to them. Clear advice on technical capability is required so they can make informed choices. Best practice should be evangelised, so that researchers can develop their skills and own practice. Points of contact for advice and support need to be established, and made public.

6.1.2 Medium-term

The medium term (1-3 years) presents opportunities to enhance research capability and profile:

Research Information Management Framework

A coherent research information management framework should be developed to integrate policy, governance, implementation and research operation. This could extend beyond research data iteself. Most of this exists, with this activity focussing on bringing it together so that it is more understandable to researchers.

Comprehensive and affordable backup service for all

Data backup is critical, but providing a comprehensive service with sufficient capacity and service levels at a manageable cost is a significant challenge. Over the medium-term such a service needs to be developed, but it requires not only a technical solution, but also organisational thought as to the cost-benefits of backing up different classes of data.

Open research data mandate, and supporting infrastructure

The University of Southampton led the open access movement and is at the forefront of open data. Many researchers already openly publish their data, but it is not supported centrally. This is a tremendous opportunity to raise the profile of Southampton's research. It requires investment to realise this, although it is one of the advantages of deploying an institutional data repository that could also support this mode of data publication.

Research data lifecycle management

Comprehensive support for research data across its whole lifecycle. Significant Work needs to be carried out to support data curation and preservation, technically, organisationally, and with sufficient resource.

Embedding data management training and support

High-quality training and support in data management best practice is necessary to ensure that researchers take advantage of the investments made by the institution.

6.1.3 Long-term

Long-term aspirations can provide significant benefits realisation across the whole University, and a stable foundation for the future:

Provide coherent data management support across all disciplines

It is envisaged that on a three year timescale it may be possible to fully support researchers, with supply meeting demand via an easy-to-use data management service. This would significantly improve research productivity, allowing them to concentrate on their research, rather than worrying about data management logistics.

Embed exemplary data management practice across the institution

If short- and medium-term recommendations are successfully implemented, then raising the baseline of data management across the institution should be possible. Researchers should feel enabled to do their research, rather than held back by current constraints and difficulties with their data management.

Agile business plan for continual improvement

Having spent three years raising the baseline capability and implementing a business plan, it will need to continually evolve as requirements and technology changes. For example, new business models such as Cloud Computing, could change the landscape significantly. The University needs to be agile in order to take advantage of new opportunities as they arise.

7 References

1. University of Southampton, 2010, “Research Integrity and Academic Conduct”, viewed at: http://www.southampton.ac.uk/ris/policies/integrity.html

2. Research and Innovation Services at UoS, 2010, viewed at: http://www.southampton.ac.uk/ris/commercialisation/index.shtml

3. Jones S, 2009, “A report on the range of polices required for and related to digital curation”, DCC policies report, viewed at:http://www.dcc.ac.uk/sites/default/files/documents/reports/DCC_Curation_Policies_Report.pdf

4. RIN, 2008, “Stewardship of digital research data: a framework of principles and guidelines”, viewed at: http://www.rin.ac.uk/files/Research%20Data%20Principles%20and%20Guidelines%20full%20version

5. Donnelly M and Jones S, 2010, “Checklist for a data management plan”, DCC report, viewed at http://www.dcc.ac.uk/resources/data-management-plans

6. EC Document, 2010, “Open access pilot in FP7, Viewed at: ftp://ftp.cordis.europa.eu/pub/fp7/docs/open-access-pilot_en.pdf

7. Hickman L, 2009, "Climate sceptics claim leaked emails are evidence of collusion among scientists". The Guardian Newspaper, London). http://www.guardian.co.uk/environment/2009/nov/20/climate-sceptics-hackers-leaked-emails.

8. Revkin A, 2009, "Hacked E-Mail Is New Fodder for Climate Dispute", New York Times, USA, viewed at: http://www.nytimes.com/2009/11/21/science/earth/21climate.html.

9. House of Commons Science and Technology Committee Report, 2010, "The disclosure of climate data from the Climatic Research Unit at the University of East Anglia".

10. Oxburgh R, Davies H, Emanuel K, Graumlich L, Hand D, Huppert H, Kelly M, 2010, "Report of the International Panel set up by the University of East Anglia to examine the research of the Climatic Research Unit" , viewed at http://www.uea.ac.uk/mac/comm/media/press/CRUstatements/SAP. "

11. UCISA, 2010, “Security Edition 3.0” viewed at: http://www.ucisa.ac.uk/publications/toolkit/ist_sections.aspx .

12. Freedom of Information Ltd., 2010, viewed at: http://www.freedomofinformation.co.uk/content/view/8/26/.

13. BBC News, 2010, viewed at: http://news.bbc.co.uk/1/hi/northern_ireland/8623417.stm.

14. Keenan D, 2010, viewed at: http://www.informath.org/apprise/a3900.htm.

15. UoS Corporate Services, 2010, “Data Protection and Freedom of Information”, viewed at: http://www.soton.ac.uk/corporateservices/legalservices/practiceareas/dpfoi.html.

16. Ames, Morgan and Naaman, Mor (2007). Why We Tag: Motivations for Annotation in Mobile and Online Media. *Conference on Human Factors in Computing Systems*, Apr 28-May 03, San Jose CA. 971-980

17. Angus, Emma et al (2008). General patterns of tag usage among university groups in Flickr. *Online Information Review* 32(1) 89-101

18. Cox, A.M. (2008). Flickr: a first look at user behaviour in the context of photography as serious leisure. *Information Research* 13(1) 336

19. Ding, Ying (2009). Perspectives on Social Tagging. *Journal Of The American Society For Information Science And Technology* 60(12):2388–2401

20. Heckner, Markus et al (2009). Tagging Tagging. Analysing User Keywords in Scientific Bibliography Management Systems. *JoDI - Journal of Digital Information* 9(2) 19

21. Hollink, L et al (2004). Classification of user image descriptions. *International Journal of Human-Computer Studies* 61(5) 601-626

22. ICAT Project (2010) http://code.google.com/p/icatproject/wiki/IcatMain

23. IUPAC (2010). International Chemical Identifier (InChi), International Union of Pure and Applied Chemistry (IUPAC), http://www.iupac.org/inchi/

24. JISC (2006). *Choosing a File Format for Digital Still Images*. http://www.jiscdigitalmedia.ac.uk/stillimages/advice/choosing-a-file-format-for-digital-still-images/

25. JISC (2009). *Systems for Managing Digital Media Collections* http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/systems-for-managing-digital-media-collections/

26. JISC (2010a). *Approaches to Describing Images*. http://www.jiscdigitalmedia.ac.uk/stillimages/advice/approaches-to-describing-images/

27. JISC (2010b) *Metadata Standards and Interoperability* http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/metadata-standards-and-interoperability/

28. JISC (2010c) *Controlling your Language: a Directory of Metadata Vocabularies* http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/controlling-your-language-links-to-metadata-vocabularies/

29. Koch, Traugott (2005). *eBank-UK Metadata Application Profile*. http://www.ukoln.ac.uk/projects/ebank-uk/schemas/profile/

30. B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman, LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision, pages 157-173, Volume 77, Numbers 1-3, May, 2008.

31. Matthews B et al (2009) Using a Core Scientific Metadata Model in Large-Scale Facilities, *5th International Digital Curation Conference (IDCC 2009)*, London, UK, 02-04 Dec 2009, http://epubs.cclrc.ac.uk/work-details?w=51838

32. NISO. (2004) "*Understanding Metadata*". NISO Press. http://www.niso.org/publications/press/UnderstandingMetadata.pdf

33. Panofsky, E. (1962). Studies in iconology. Harper & Row, New York.

34. Patel, Manjula (2010) *I2S2 D1.1 Requirements Report*, http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2-WP1-D1.1-RR-Final-100707.pdf

35. Pavlidis, Theo (2009). Why Meaningful Automatic Tagging Of Images Is Very Hard. *IEEE International Conference On Multimedia And Expo*, 1-3, 1432-1435

36. Scott, M (2010). Materials Data Centre, *PhD Nine Month Report, School of Engineering Sciences, University of Southampton* [unpublished]

37. Shatford Layne, S. (1994). Some issues in the indexing of images. *Journal of the American Society for Information Science* 45(8):583-588.

38. Shatford, S. (1986). Analyzing the subject of a picture: a theoretical approach. *Cataloging & Classification Quarterly* 6(3):39-62.

Appendix I - Questionnaire

Institutional Data Management Blueprint Survey

Participant consent form

Please read the following carefully before agreeing to take part in this study; I understand that:

- All results from this study will be anonymous. Information extracted from this questionnaire and any subsequent interview will not, under any circumstances, contain names or identifying characteristics of participants:
- I am free to withdraw from this study at any time without penalty:
- I am free to decline to answer particular questions:
- Whether I participate or not there will be no effect on my progress in employment in any way:

Section 1. About you

Question 1.1

Please enter your full name.

Question 1.2

Please enter your School.

- Chemistry
- Engineering Sciences
- Electronics and Computer Science
- Humanities
- Other

Question 1.2b

Please give details

Question 1.3

Please describe your research role.

Academic or equivalent (e.g. HEFCE-funded)	Research Fellow (e.g. Project-funded)	Research Support (e.g. Technician)	Research Student	Other
○	○	○	○	○

Question 1.3b

Please give details.

Question 1.4

What is you area of research?

Archaeology	Chemistry	Materials	Nano Science	Computational modelling	Other
○	○	○	○	○	○

Question 1.4b

Please give details.

Question 1.5

Do you currently hold/store any research data?

- ○ Yes
- ○ No

Section 2. About your data

Question 2.1

Who do you believe owns your research data?

Yourself	Project Team	School	Institution	Funding body	Don't know	Other
○	○	○	○	○	○	○

Question 2.1b

If you answered 'Funding Body give details below

Question 2.1c

If you answered 'other' please give details below

Question 2.2

Characteristics of the data – please tick all that apply

- ☐ Observational (of specific phenomena at a specific time or location where the data will usually constitute a unique and irreplaceable record)
- ☐ Experimental (scientific experiments and computational results, which may in principle be reproduced although it may in practice prove difficult or not cost-effective to do so)
- ☐ Computer code (including model or simulation source code, where it may be more important to preserve the model and associated metadata than the computational data arising from the model)
- ☐ Reference (canonical or reference data relating for example to gene sequences, chemical structures or literary texts)
- ☐ Derived (resulting from processing or combining "raw" or other data (where care may be required to respect the rights of the owners of the raw data))
- ☐ Other - please give details

Question 2.2b

If you answered 'other' please give details below

Question 2.3

Data types – please tick all that apply

- ☐ Data automatically generated from or by computer programs
- ☐ Data collected from sensors or instruments (including questionnaires)
- ☐ Computer software source code
- ☐ Images scans photos or X-rays
- ☐ Websites
- ☐ MS Word or equivalent word processing software
- ☐ Excel sheets or equivalent spreadsheet software
- ☐ MS PowerPoint or equivalent presentation software
- ☐ SPSS files or equivalent statistical software
- ☐ MS Access or equivalent database software
- ☐ Digital audio files
- ☐ Digital video files
- ☐ Fieldwork data
- ☐ Laboratory notes
- ☐ Video tapes
- ☐ Audio tapes
- ☐ Slides - physical media
- ☐ Other - please give details

Question 2.3b

If you answered 'other' please give details below

Question 2.4

Do you currently have any data which is no longer compatible with existing software or on hardware media that are not now widely readable?

- Yes
- No

Question 2.4b

If yes, please give details (age, type of software/hardware, etc.)

Question 2.5

Have you ever re-used your own data from previous projects?

- Yes
- No

Question 2.5b

If yes, what are your experiences of locating and using such data?

Question 2.6

Have you ever used data from external sources?

- Yes
- No

Question 2.6b

If yes, what are your experiences of locating and using such data, costs, licenses etc?

Question 2.7

Who is responsible for managing your data? - please tick all that apply

- ☐ Project manager
- ☐ You
- ☐ Research groups
- ☐ Research Assistant
- ☐ Nobody
- ☐ National data centre or data archive - give details
- ☐ Other - give details

Question 2.7b

If you answered 'National data centre or data archive' please give details below

Question 2.7c

If you answered 'Other' please give details below

Question 2.8

Where do you store your current data? - please tick all that apply

- ☐ Paper/file records
- ☐ Local computer
- ☐ CD/DVD
- ☐ USB/Flash Drive/Memory stick
- ☐ External Hard Disk
- ☐ My documents on iSolutions PC
- ☐ iSolutions provided file server
- ☐ External/commercial/web data storage facility - give details
- ☐ Other provided file server e.g. by School/unit - please specify
- ☐ Other - give details

Question 2.8b

If you answered 'External/commercial/web data storage facility' please give details below

Question 2.8c

If you answered 'Other provided file server e.g. by School/unit' please give details below

Question 2.8d

If you answered 'Other' please give details below

Question 2.9

How much electronic data do you currently retain? – please tick all that apply

	N/A Paper/film based data	Up to 2 Gigabytes (GB) (up to one writeable DVD full)	2 - 20 GB	20 - 100 GB	100 - 500 GB	500 GB - 1 Terabyte (TB 1000 GB)	1 TB or more	Don't know
Current project	○	○	○	○	○	○	○	○
In total	○	○	○	○	○	○	○	○

Question 2.10

How long do you keep your data?

- ○ Until the end of a project/body of work/when results are published
- ○ Up to a year after the end of the work
- ○ 1 - 5 years
- ○ 5 - 10 years
- ○ 10 - 25 years
- ○ 25 - 50 years
- ○ 50 years or more but with a defined lifetime
- ○ Forever
- ○ Don't know

Question 2.11

Do you keep data for compliance reasons?

- Yes
- No

Question 2.11b

How long do you keep it?

Question 2.12

Have you ever experienced storage problems due to the size of the files?

Yes	No

Question 2.12b

How did you overcome these storage issues? (please tick all that apply)

Requested additional storage space from iSolutions	Purchased an external hard drive	Saved to portable media e.g. memory stick/USB/flashdrive/CD	Other

Question 2.12c

If you answered 'other' - please give details below

Question 2.13

How frequently do you backup your data?

- I do not back up my data
- No fixed schedule - when I remember
- At the end of a project/body of work

- ○ At least annually
- ○ At least quarterly
- ○ At least monthly
- ○ After every update
- ○ Automatically via iSolutions nightly backup (retained for three months)

Question 2.14

Where do you back up your data? - please tick all that apply

- ☐ Paper/file records via photocopy of similar
- ☐ Another computer
- ☐ CD/DVD
- ☐ USB/Flash Drive/Memory stick
- ☐ External hard disk
- ☐ My own tape backup system
- ☐ School/unit - provided file server
- ☐ My documents on iSolutions PC
- ☐ To iSolutions provided file server
- ☐ To iSolutions backup system
- ☐ To external/commercial/web data storage facility - give details
- ☐ Other - give details

30471

Question 2.14b

If you answered 'External/commercial/web data storage facility' please give details below

Question 2.14c

If you answered 'Other' please give details below

Question 2.15

Do you deposit your data with other services, such as the UK Data Archive?

- ○ Yes
- ○ No

Question 2.15b

If yes please give details

Question 2.16

How do you keep track of where your data is stored, and what it relates to? - please tick all that apply

- ☐ In a paper logbook
- ☐ In an electronic logbook
- ☐ In a spreadsheet
- ☐ In a local database (e.g. research group)
- ☐ In a remote database (e.g. iSolutions national archive)
- ☐ Other - please give details

Question 2.16b

If you answered 'Other' please give details.

Question 2.17

During the lifetime of a project, do you allow others access to data on which you are working?

- ○ Yes
- ○ No

Question 2.17b

If yes give details e.g. external users; partners in collaboration etc

Question 2.17c

If no, what access issues are of concern to you? Please tick all that apply

- ☐ Confidentiality or data protection issues
- ☐ Licence agreements prohibiting sharing
- ☐ The data is not fully documented
- ☐ The data is no longer is a format that is widely readable/accessible
- ☐ Sharing not required
- ☐ Other - give details

Question 2.17d

If you answered 'other' please give details below

Question 2.18

Do you allow others to access your data once the project is finished?

- ○ Yes
- ○ No

Question 2.18b

If yes give details, e.g. others in research group, external users, project partners', etc

Question 2.18c

If no, what access issues are of concern to you? Please tick all that apply

- ☐ Confidentiality or data protection issues
- ☐ Licence agreements prohibiting sharing
- ☐ The data is not fully documented
- ☐ The data is no longer is a format that is widely readable/accessible
- ☐ Other - give details

Question 2.18d

If you answered 'other' please give details below

Question 2.19

Have you ever been asked by a funder to produce a Data Management Plan?

- Yes
- No

Question 2.19b

Please give details

Question 2.20

Are there any data preservation policies in place within your School e.g. data preservation policy, record management policy or data disposal policy?

- Yes
- No
- Don't know

Question 2.20b

If yes, please give details

Question 2.21

Would you find it useful to have university wide guidelines to manage and maintain your research data?

- Yes
- No

Question 2.22

What is the biggest problem for you with regard to managing and storing your data?

Question 2.23

How can the University (including your School, iSolutions and the Library) make data management and storage easier for you?

Question 2.24

Please confirm if you would be willing to participate in a short follow-up interview to this survey (max 1 hr)

- Yes
- No

Question 2.24b

Please enter your email address

Section 3. Final comments

Question 3.1

Do you have any further comments you would like to make?

Many thanks for taking the time to complete this questionnaire. (You can keep track of this project by checking http://www.southamptondata.org.)

Appendix II - Interview Questions

Managing data

A variety of methods of managing/preserving data are used throughout the university.

1. Do you consider preservation of data worthwhile?

2. What do you think are the advantages / disadvantages?

3. How far do you feel supported with carrying out data management?
   - What support is available? Is it sufficient for your needs?
   - How would you like data management support to look in an ideal world?

4. How can the University help you manage and maintain your research data?

5. What are the major issues you are concerned about?

6. How do you label, annotate and organise your data?

7. Are there any standards for this in your area, that you use or are aware of?

Storage

1. Do you store all your data for a project in the one place, or is it distributed amongst various media/places/people?

2. Do you currently store physical data – i.e. maps, slides, paper files, etc. If so how is this managed?

3. How can someone find out about the data you hold – is it possible to find it easily?

4. How do you ensure long term security of any physical hardware on which data is stored?

5. Do you consider the digital information you hold to be safe?

6. Why did you choose to store your data in the way you have? Were your choices limited in any way by the anticipated lifespan, physical space, importance or confidentiality?

7. Have you ever lost any data, and if so, how did this happen?

8. How do you plan to store digital data in the future?

9. Have you managed to overcome any storage problems due to the size/number of the files, and if so – how?
   Did this problem change the way you worked?

10. Have you ever asked for advice about storage – and if so, was it helpful/successful? Who provided helpful advice?

Compatibility

1. If you have had a problem with the compatibility of your data with current hardware and software, did you manage to find a solution? Do you use hardware that is no longer readily supported?

2. If you have data stored externally as well as in house, are there any compatibility issues?

Schools policies / Data management plan

Increasingly funding bodies want researchers to include a data management plan in the funding application.

1. If you have had to do this:
   - Which funder?
   - How did you find this experience?
   - Did it affect how you shared/managed your data?

2. If you have NOT had to do this:
   - Do you do any planning before you use or collect data; what have you done?

3. Is there any data management guidance in areas such as data preservation; record management or data disposal in place within your school or even your research group?

   - If so, what policies are in place?
   - How useful are they?
   - Would it be useful to have a university wide policy in place? What would be your major concerns?

Collaboration/Sharing

1. Have you ever shared your data, either on request or in the context of collaboration?

- Who with? Where were they geographically? How many of you were sharing
- What did you do? What methods did you use to share?

- What did or didn't work when you were sharing data and how did you deal with; version control, file naming conventions, legal issues transferring data?
- How was ownership decided?
- Have you ever shared confidential data? How did you do it?

2. What is your experience of the availability, quality and usefulness of this sort of data?

3. Would you have any reservations about sharing your data?

4. To non-collaborators
   - How do you deal with the day-to-day managing of data?
   - Version control
   - Labelling files etc (as above)

Fiona Nichols

Updated 25 January 2010

Appendix III – AIDA

AIDA for Research Data: Departmental Assessment	Resources				
	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**
	ACKNOWLEDGE / NO ACTION	**ACT / LOCALISED**	**CONSOLIDATE / CO-OPERATE**	**UNIFY DEPARTMENT / INTERNAL INTEGRATION**	**EXTERNALISE / EMBED**
	Very low confidence. *Nobody in the Department is doing this.* *We don't have any formalised financial or staffing policies.* *We do not have financial autonomy.* *We have no evidence of any action.* *These things are implied rather than actually carried out.* *We don't meet the benchmark but we acknowledge this is the case.*	*Some confidence.* *At least one person in the Department is doing this.* *We have evidence of some local activity.* *Practices can vary, and are ad-hoc and inconsistent.* *Work on this is still unfinished or it has only just started.* *Financial allocation in this area is uneven.*	*Medium confidence.* *At least three people in the Department are doing this, and are doing so in harmony with each other.* *Practices are consistent within the Department.* *Their activities still cover only a few defined areas of managing research data, not everything.* *These actions are local and only affect our Department.* *We are not yet harmonised with the entire Institution.*	*High confidence.* *Everyone in our Department / Research Group is doing this.* *These actions are fully in place.* *All defined areas of financial practice, funding and staffing are covered.* *We have a strong evidence base and can demonstrate these things.* *However, although we're all integrated and harmonised, the rest of the University hasn't caught up with us yet.*	*Very high confidence.* *Everyone in our Department / Research Group is doing this, and it is embedded in our workflow to the point we don't have to think about it.* *All new staff who join the group comply with this.* *No staff members are left out or overlooked.* *We are harmonised with the rest of the Institution.* *We may be working, where appropriate, to agreed external standards.* *We are working, where appropriate, with others outside the University.*
RES 01: Financial sustainability plan					
The business plan supports the sustainability of our research data					
We generate income through our research data					
Our research data collection is self-supporting					
RES 02: Review of business plan					
RES 03: Technological resources allocation					
Sufficient money is being invested on the technology we need for our research data (not just storage)					
There are dedicated funds available for technology development in support of our research data					
Technology Watch is in place for emerging technologies					
Future technological requirements are anticipated					
Department is capable of assigning the necessary technological resources to the research data collection					
RES 04: Risk analysis					
We have a formal risk management plan in case of data loss					
Risk analysis is based on existing standards					
RES 05: Transparency and auditability					
RES 06: Sustainability of funding for research data					
There will be enough money to keep our research data safe					
Funding is inbuilt to the core function of our Department					
RES 07: Staff skills					
Department has the requisite skills available to manage its research data					
Our funding enables the steady maintenance of core staff skills					
RES 08: Staff numbers					
Department has enough staff to manage its data					
RES 09: Staff development					
Staff are competent in research data management					
Staff skillsets have currency					
Staff skillsets evolve in line with technological changes					
We have a professional development and training policy					
We have a training budget					

Appendix IV – Funders' Policies

Curation policies and support services of the main UK research funders

	Policy coverage		Policy stipulations					Support provided		
Research Funders	Published outputs	Data	Time limits	Data plan	Access / sharing	Long-term curation	Monitoring	Guidance	Repository	Data centre
AHRC - Arts and Humanities Research Council	●	●	●	●	●	◐	○	●	○	◐
BBSRC - Biotechnology & Biological Sciences Research Council	●	●	●	●	●	●	●	◐	●	○
EPSRC - Engineering and Physical Sciences Research Council	●	○	●	○	●	○	○	○	○	○
ESRC - Economic and Social Research Council	●	●	●	●	●	●	●	●	●	●
MRC - Medical Research Council	●	●	●	●	●	●	○	●	●	◐
NERC - Natural Environment Research Council	●	●	●	●	●	●	●	●	●	●
STFC - Science and Technology Facilities Council	●	○	●	○	●	○	○	○	●	◐
Wellcome Trust	●	●	●	●	●	●	●	○	●	○

Terminology clarifications

Published outputs: a policy on published outputs e.g. journal articles and conference papers
Data: a data policy or statement on access to and maintenance of electronic resources
Time limits: set timeframes for making content accessible or preserving research outputs
Data plan: requirement to consider data creation, management or sharing in the application
Access / sharing: promotion of OA journals, deposit in repositories, data sharing or reuse
Monitoring: whether compliance is monitored or action taken such as withholding funds
Curation: stipulations on long-term maintenance and preservation of research outputs
Guidance: best practice guides or curation support staff available to funded researchers
Repository: provision of a repository to make published research outputs accessible
Data centre: provision of a data centre to curate unpublished electronic resources or data

KEY:

● full coverage

◐ partial coverage

○ no coverage

Published outputs

The research funders' policies on published outputs are aligned with the joint RCUK position statement, which was first issued in June 2005.[26] All advocate open access to outputs from their funded research programmes and many provide a repository service in support of this requirement. There are differences regarding how publications fees should be met, and some funders include additional stipulations – NERC, for example, may take compliance with this policy into account when considering further applications for funding.

Data

Most funders have some form of policy regarding data, however the extent and coverage of these vary greatly. In several cases researchers are directed to good practice guides, which provide recommendations on documenting and maintaining research. There are only two Research Councils without a formal data policy as yet – the EPSRC and STFC – though a data policy is currently being developed at EPSRC and STFC appears to continue CCLRC and PPARC procedures.

Time limits

The timeframes stipulated for access and curation vary by funding body. Most expect publications to be made openly available as soon as possible or in a timely manner, which is generally understood to be at least within six months of publication of results. The ESRC and AHDS (only in the case of archaeology) expect an offer of deposit of data within three months of the end of the award, and also advocate a relationship with the data centre from the outset of the project. The BBSRC, MRC and Wellcome Trust have a general statement that data must be kept securely for a period of ten years after the completion of a research project, while the EPSRC maintains it should be held for an 'appropriate' length of time.

Data plan

Most research funders require applicants to submit a statement on access, management and long-term curation of their research outputs at the proposal stage. The focus of this statement varies by funder: the AHRC, ESRC and NERC all require a statement on how resources will be created so they can be preserved in the long-term, while the BBSRC, MRC and Wellcome Trust focus heavily on the data sharing potential of research resources. Neither the EPSRC nor STFC currently require applicants to submit data sharing or curation plans as part of the proposal, however this is likely to change as data policies emerge.

Access / data sharing

All funders have signed up to the RCUK statement on open access of research outputs and advocate making publications widely accessible. They largely agree to meet publication fees, normally as indirect costs, to ensure research is freely accessible. The MRC and Wellcome Trust also encourage - or in cases where they have paid publication fees, require - licences that allow articles to be freely copied and reused for purposes such as text and data mining. Some moves are also being made towards linking publications with source data, for example UK PubMed Central[27] allow deposit of supplemental material in support of the publication.

The concept of open data is not advocated in any of the research funders' policies, however the BBSRC, MRC and Wellcome Trust have the strongest ethos of data sharing, expecting data to be made available with as few restrictions as possible. The ESRC and NERC facilitate data sharing through their funded data centres, however licence fees and access restrictions are often applied as their remit is to serve research and teaching communities. The AHRC provides access to, and a cross-search of, their funded archaeology data through the ADS and requires other award holders to keep data accessible for a minimum of three years. The STFC has not provided a clear statement on expectations for data sharing. According to a

[26] See RCUK website at: http://www.rcuk.ac.uk/access/default.htm

[27] See website at: http://ukpmc.ac.uk/

RIN study the EPSRC considers the discipline areas it covers do not have so much need for data sharing,[28] however the policy that is currently being developed may revise this position.

Long-term curation

Most of the funders consider curation in detail in their policies. The AHRC, BBSRC, ESRC, MRC, NERC and Wellcome Trust all consider various aspects of the curation lifecycle, for example noting the need to create resources according to appropriate standards and best practice, maintain adequate documentation and metadata to ensure usability, and manage data appropriately in the short-term so it can be preserved for the future. The EPSRC only has one stipulation – that data be appropriately stored for a minimum of 10 years – while the STFC does not appear to have any formal requirements addressing curation at present.

Monitoring

NERC and the Wellcome Trust note they monitor compliance with the open access policy on publications. The BBSRC will monitor adherence to the data management and sharing plan and may take this into consideration for future proposals, while the ESRC could withhold the final grant payment if data is not deposited on time. The extent to which such penalties are applied is unclear. The other funders meanwhile do not appear to monitor adherence or impose penalties for non-compliance with their curation policies.

Guidance

The extent to which guidance and support services are provided varies significantly. The best served researchers are those funded by the ESRC, which provides extensive curation guidance through the UK Data Archive arm of the ESDS,[29] and NERC, whose data centre staff will provide assistance and advice throughout the award.[30] The AHRC runs a similar service for archaeology researchers and has legacy guides online for researchers in other fields.[31] The MRC meanwhile is setting up a data support service and already provides some best practice guides and data sharing toolkits. BBSRC does not appear to have much guidance online, but states that information on relevant standards and best practice will be provided and a main contact for this is listed. No particular sources of guidance were noted by the EPSRC, STFC and Wellcome Trust in their policies or found on their websites. It may be that curation support is offered less formally by these funding bodies.

Repository

Most research funders provide a publications repository for their funded researchers. ESRC, NERC and STFC all run their own e-Prints service while BBSRC, the MRC and Wellcome Trust are partners in PubMed Central. The only Councils that do not provide a repository for published outputs are the AHRC and EPSRC. Researchers supported by these Councils are expected to use any institutional or subject based repositories available to them.

Data Centre

Provision of data centres is patchy - very few funding bodies have a full data service in place. The exceptions are the ESRC and NERC, which both provide comprehensive preservation and support services. The BBSRC, MRC and Wellcome Trust meanwhile agree the cost of long-term curation can be included in the original proposal. The AHRC provides a data service for researchers in the area of archaeology through ADS and it appears STFC have several services and agreements in place to provide pockets of support, for example through the UK Solar System Data Centre. The Wellcome Trust, BBSRC, MRC and EPSRC all contribute to the European Bioinformatics Institute,[32] however for research that falls outside the EBI remit the institutions in which funded researchers are based are expected to maintain outputs in the long-term.

---

[28] RIN, *Research funders' policies for the management of information outputs,* (2007), p61, available at: http://www.rin.ac.uk/files/Funders%27%20Policy%20&%20Practice%20-%20Final%20Report.pdf

[29] For the data management and sharing guidance see: http://www.data-archive.ac.uk/sharing/

[30] For details of the data centres see: http://www.nerc.ac.uk/research/sites/data/

[31] Details of the ADS are at: http://ads.ahds.ac.uk/ and other legacy guidance is available from the archived AHDS website at: http://www.ahds.ac.uk/about/publications/index.htm

[32] For more information on EBI, see: http://www.ebi.ac.uk/