



## JISC Final Report


Project Information			
<b>Project Identifier</b>	<i>To be completed by JISC</i>		
<b>Project Title</b>	Institutional Data Management Blueprint		
<b>Project Hashtag</b>			
<b>Start Date</b>	1 Oct 2009	<b>End Date</b>	Extended to 31 Aug 2011
<b>Lead Institution</b>	University of Southampton		
<b>Project Director</b>	Kenji Takeda (until May 2011); Jeremy Frey (from June 2011)		
<b>Project Manager</b>	Clint Styles		
<b>Contact email</b>	<a href="mailto:w.h.white@soton.ac.uk">w.h.white@soton.ac.uk</a> or <a href="mailto:m.l.brown@soton.ac.uk">m.l.brown@soton.ac.uk</a>		
<b>Partner Institutions</b>	Digital Curation Centre, University of Oxford, National Oceanography Centre		
<b>Project Web URL</b>	<a href="http://www.southamptondata.org">http://www.southamptondata.org</a>		
<b>Programme Name</b>	Managing Research Data (Research Data Management Infrastructure)		
<b>Programme Manager</b>	Simon Hodson		

Document Information			
<b>Author(s)</b>	Mark Brown & Wendy White		
<b>Project Role(s)</b>	Members of IDMB Project Team		
<b>Date</b>	31.8.2011	<b>Filename</b>	IDMB_finalreport.docx
<b>URL</b>	<i>If this report is on your project web site [will be posted after acceptance]</i>		
<b>Access</b>	This report is for general dissemination		

Document History		
Version	Date	Comments
1	5.8.2011	MLB initial draft
2	18.8.2011	WHW edit
3	30.8.2011	MLB/WHW edit

## Table of Contents

NB : This table of contents 'auto-populates' - to update the table of contents – place cursor in the table of contents, right-click your mouse, click 'update field', select appropriate option

	1
<b>1 PROJECT SUMMARY</b>	<b>3</b>
<b>2 MAIN BODY OF REPORT</b>	<b>4</b>
2.1 PROJECT OUTPUTS AND OUTCOMES	4
<i>Data management questionnaire results</i>	4
2.2 HOW DID YOU GO ABOUT ACHIEVING YOUR OUTPUTS / OUTCOMES?	4
2.3 WHAT DID YOU LEARN?	6
2.4 IMMEDIATE IMPACT	7
2.5 FUTURE IMPACT	7
<b>3 CONCLUSIONS</b>	<b>8</b>
<b>4 RECOMMENDATIONS</b>	<b>9</b>
<b>5 IMPLICATIONS FOR THE FUTURE</b>	<b>9</b>

## 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 has been run as a partnership between staff from across the University who have contributed a range of knowledge and skill.

The project investigators: Kenji Takeda (Engineering Sciences, Project Lead); Mark Brown (University Librarian); Simon Coles (Chemistry); Les Carr (Electronics & Computer Science); Graeme Earl (Archaeology); Jeremy Frey (Chemistry); Peter Hancock (iSolutions); Wendy White (Library: Head of Scholarly Communication); Clint Styles (Project Manager).

The investigators have been a strong team from iSolutions (Steve Patterson) and the Library: Fiona Nichols; Michael Whitton; Harry Gibbs; Christine Fowler and Pam Wake.

The project team would like to thank the internal and external members of 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), and in addition the input from the Archaeology Data Centre and the AIDA team which has been invaluable. Advice and support from the DCC has been particularly appreciated.

We would also like to acknowledge the contribution made by Simon Hodson as Programme Manager at JISC

## 1 Project Summary

Research data management has become an international topic of concern to researchers and their funders. Defining the responsibilities for managing research data from inception to preservation is recognised as a complex process shared between individual researchers and research groups, institutions, funders and national agencies. To respond to these needs institutions require an overall framework within which to plan and develop their data management strategy. The Institutional Data Management Blueprint Project has worked with senior managers, researchers and services across the University to define a potential framework, identify immediate priorities and highlight the key elements to be achieved over a ten year period.

This Blueprint combines two approaches. A bottom up approach based on researchers' needs, designed to enable them to adopt good practice, and a top-down approach, intended to provide the institutional policies and infrastructure to be effective. The project used a research-led approach derived from an in-depth survey of the needs of researchers in representative disciplines the use of the AIDA audit tool and focused work with exemplar disciplines, particularly Archaeology. This showed that despite much good practice shared locally through local networks, institutionally there was no overall strategy for defining roles or for supporting researchers.

In response to this the project has defined a draft data management policy which will form part of wider work on a set of information management principles. It has set out an investment strategy with detailed business plan for advanced data storage, and a core metadata structure for data deposit. Responding to researchers' concerns with the overhead and complexity of managing data deposit, the project has initiated a research data repository on Eprints, offering the opportunity to integrate data management with research outputs and publication in line with the University's commitment to open data and open access. Initial work was also done on assessing the potential to Sharepoint to take forward the repository role in due course.

The audits reiterated that researchers are best incentivised to take up new practices and processes by having a clear policy and service support. The Blueprint defines a multifunctional team approach which can bring together the knowledge and expertise of both researchers and professionals within an institutional policy and technical framework for managing their data.

The Blueprint, and the ten year roadmap aim to meet the needs of a multi-disciplinary, research-intensive University, and is designed to be both practical and iterative.

## 2 Main Body of Report

### 2.1 Project Outputs and Outcomes

<b>Output / Outcome Type</b> <i>(e.g. report, publication, software, knowledge built)</i>	<b>Brief Description and URLs (where applicable)</b>
Institutional Data Management Blueprint	Blueprint document outlines framework model and includes 10 year Roadmap
IDBM Initial Findings Report	Outputs from the baseline survey of current practice and researchers' views at the University of Southampton <a href="http://www.southamptondata.org/idmb-blog.html">http://www.southamptondata.org/idmb-blog.html</a>
Data management questionnaire results	Outputs from the extended survey on researchers' data management practice at the University of Southampton
Archaeology Case Study	Covers exemplar including information on research data practice data-gathering and training pilot.
Research Data Management Policy	Appendix A of Blueprint
Business Model for Data Storage	Appendix B of Blueprint
Presentations:	<a href="http://www.southamptondata.org/idmb-blog.html">http://www.southamptondata.org/idmb-blog.html</a>
Eduserv Symposium 18/5/2011	
6 <sup>th</sup> DCC Conference 8/12/2010	

### 2.2 How did you go about achieving your outputs / outcomes?

This was a deliberately ambitious project. The Project Team were aware that within the University there was a considerable amount of very developed practice in terms of data management and the aim was to draw on this good practice and extend it across the institution but particularly to those disciplines where practice was not so developed. From the start therefore the project was conceived as a multidisciplinary institutional project which would take a high-level view of what the institution itself could achieve in providing a more coherent framework for data management itself.

The University of Southampton is a multidisciplinary research environment with extensive national and international links. As one of the centres for e-science, and with strong links to some of the national data centres including the archaeology data centre, ESRC data management model and NERC data centre as well as locally managed material in the areas of electronics and computer science and chemistry. Southampton has also taken part in debates at national level about achieving a balance between national institutional framework for data management and those elements of the data management process required to be managed institutionally, most specifically in the debate surrounding the UKRDS. The debates within RLUK and RUGIT on the importance of computing and library input into the data management lifecycle also prompted consideration of academic and service partnership as a means of achieving a way forward at institutional level for the effective management of research data. In the case of the University of Southampton the institutional commitment to open access, by extension open data, and the extensive work with repositories made us believe that this

was an excellent opportunity to develop a more coherent and cohesive institutional framework for data management.

In scoping the project we had a number of key aims:

- to create a practical and attainable institutional framework for managing research data throughout its lifecycle that could encompass the whole institution;
- to develop the key elements of this framework through analysis of current data management requirements for a representative group of disciplines with a range of different data;
- to build upon the range of current work on research data management to inform the framework and draw out practical applications;
- to determine the role of institutional policy in matching the range of current and potentially future requirements by funders for data management;
- to gather informational current practice and perceived priorities within the academic community to match the framework with actual needs;
- to create within the research community at Southampton an appreciation of the value to data management including both good and data management practice and the value of sharing data;
- to develop a service model based on the close collaboration between researchers and service providers drawing together technical, organisational and professional expertise from across the institution;
- to establish a realistic and obtainable metadata strategy that could ensure an effective registry of the University's research data assets;
- to provide a realistic business model which could be used both to develop the service infrastructure;
- to provide the basis for assessing the resource requirements for the University to meet its data management responsibilities.

This was therefore as much a cultural as a technical project, focused on the way in which an institution could respond to the challenges of increased data flow, its management and its value across the academic community. In line with this cultural perspective we wanted to create a pathfinder project for an institutional data management strategy for the next decade through a data management institutional blueprint based on these objectives support by a service orientated, extensible enterprise architecture model and a ten year business model roadmap.

Our starting point was our research community itself, to draw together a representative group from the research community and engage them through a workshop in discussing the current issues facing researchers in terms of data management. The hook for this activity was the well-known controversy over the release of climate data from the University of East Anglia and the University's inability to act coherently when faced with freedom of information requests. The workshop acted as a prelude for a wide investigation through an extensive questionnaire led survey of researchers, including a number of detailed follow-up interviews, to draw out from the community the key aspects of their concerns and their understanding of the data management landscape. Evidence from this survey was then triangulated with use of the AIDA audit methodology that investigated both individual researcher's approach to data management and the preparedness of the institution at both central and local level to respond to data management demands.

In order further to reference the conclusions from the surveys we engaged through our steering group with senior academics in the institution, including the PVC for Research, and representatives from external partnering organisations, including the DCC, the University of Oxford and the NERC data centre. This allowed us to draw out the potential implications institutionally for the data gathered from the questionnaire. As a result we extended the range of researchers covered by the questionnaires in a second round and were then able to review the policy and service provision framework that would potentially meet researchers' needs.

It was important for us to see this issue as researcher-led. We know from practical experience that researchers have a partial view of the research data lifecycle, but they are also particularly concerned about the realities of changing funder requirements, of the security and curation of data made more intense by the known losses during the recent fire at the University, and the value of being able to access and retrieve their own data, not only in support of publication but also in support of

collaborative and development work with other groups or partners. It was therefore a deliberate decision not to engage researchers with some of the highly complex, theoretical models that have provided so much valuable insight into the complexities of data management itself. Our aim was to encourage a longer term cultural chain in which the essentials for effective data management practice could be embedded across the institution and across disciplines.

It was also accepted that researchers are working in a mixed economy. In some cases funders were providing data centres for the management of that data and therefore required a different contribution from institutions to those funders, recently epitomised by the change of policy in the EPSRC, who were requiring institutions to take responsibility for the management of that data themselves. In the course of the project therefore it became obvious that researchers' immediate interest which was in the storage and security of their data while projects were active, but this concern was increasingly extending to longer term curation and preservation. Without a successful institutional approach to this effective data management policy would not be achieved.

## **2.3 What did you learn?**

In terms of understanding a researcher led approach to data management it was clear that certain key interventions supported by effective service support would be important to achieve.

Firstly, we needed to ensure that we had a data management policy which was able to appeal to researchers in terms of their needs and to provide the institution with the context for investment. Our own draft policy will go forward for ratification over the next three months and will set out policy framework supported by a number of other policy developments with which the University is currently engaged, including a data security policy and an information management policy. This emphasises the second enabler which is commitment by the institution to investing in the data management infrastructure, both in terms of its organisational, its technological and its resource implications. In a devolved and discipline diverse institution such as Southampton, it is very important to emphasise that policy are enabling frameworks rather than vehicles for compliance and achieving this balance can be challenging.

It was very important, therefore, that the third element, which was a clear commitment by the institution to invest financially in the development of the data management infrastructure, could be seen during the lifetime of this project. The heightened awareness of the value of effective data management policy for the institution and for research capability and reputation of the institution, has successfully encouraged an initial investment in data storage and data archiving resource which itself now forms part of a wider project to develop a new off-campus data centre due to open in 2013.

These first three elements are fundamentally focused at the institutional level. It is also clear from the surveys and also from the audit that the researchers themselves needed to feel individually or as a group, effective support in achieving good data management practice. The fourth element therefore was to investigate a way in which a relatively simple straight-forward metadata structure could be applied to encourage all researchers to provide core metadata for an institution-wide registry. Having worked on the core metadata elements it then became clear that a form of relatively straight-forward ingest interface was required which researchers could familiarise themselves at an institutional level regardless of discipline. The extension of the e-prints institutional research repository to include Eprints4data was a natural choice. Researchers at Southampton are already required to place their research output on the e-print repository and through the Library there is service support for the e-prints repository, including both desk-side training and direct point of need support. As an initial approach the use of the e-print software is a viable way of drawing together the community, the project was also aware that there was a danger in over reliance upon one piece of software albeit one developed and maintained by the University itself. In parallel, therefore, work that was begun with exploiting the potential of SharePoint 2010 was developed at the level of feasibility, and one of the aims of the project became to work these two options in parallel with longer term aim of migrating from one to the other should it be appropriate.

The 5<sup>th</sup> element therefore arose naturally out of this work. It is clear that researchers need a comprehensive but flexible service support system which can bring together services including the library, computing (iSolutions), research innovation services, finance and legal services. The University has a good track record of collaborative working between the services, and the data management blueprint will set out a commitment to joint working focused on data management needs.

In responding to the management of the research lifecycle, however, the Project Team is aware that some of the intricacies of the management of very large and very complex data has potentially been over simplified. Particular attention has therefore been paid to the work with archaeology data. This data covers a wide spectrum of different formats, allowing more experience to be gained at institutional level while at the same time acknowledging that the national Archaeological Data Centre provides national level expertise in this area.

The question arises for institutions as to the extent that they can provide high level understanding of data structures and data curation and preservation for data which has no national capability. Feedback from the initial work undertaken in terms of training and development with researchers by the library team reveals clearly that there will be limitations in the extent to which staff from non-cognate disciplines can provide the full range of information that researchers may require. Researchers have however shown a very strong interest in working with the services to support their needs, and this will provide an excellent basis for future collaboration.

## **2.4 Immediate Impact**

We believe that there is the basis for achieving a cultural change in the attitude of researchers to the institutional commitment to their needs. One of the most important differences is that the institution itself is now much more conscious about the need to develop the data management framework and through the data management policy, related policies, and the investment strategy put forward for the next three years, there is a real sense in which services feel that they can build upon this to develop services for the institution. Although the scenarios and pilots have been based on a limited number of disciplines, we feel that the context is now clear and the basis for future work is mapped out. This is however only the beginning. There are real issues about the extent to which the capacity of the central computing service to substitute for what has been seen as a locally managed issue and the extent to which the library and other services can provide an institution-wide service particularly in the area of metadata, standards and requirements is still very much in embryo.

We have to accept that our immediate impact has been with the senior levels of the institution including the PVC for Research, the Chief Operating Officer, the Registrar and leading academics in the pilot areas. The service heads are also fully committed to collaboration although we have yet to define the full business model and cost it accordingly. We know this because of testimony from those individuals, and because the extent to which the publicity surrounding our project has extended our reach outside the institution. Having gathered evidence about researchers' needs, the project group feels that we need to put services in place to assess their value to researchers. We need to move from the enthusiast to the everyday and this will require well trained service staff as well as advocacy and engagement with the research community as a whole.

## **2.5 Future Impact**

The main target community for future impact is the research community at the University of Southampton, as good research data management practice beds down in the policy and practice of the institution. Extended development of guidance, exemplars and support services will ensure that impact of this project is long-term. The researcher-focussed model will be further developed through iterative partnership working with IT, library and research office specialists. This will enable us to further explore the relationship between specialist needs and institutional context to ensure that support is available for the full range of required practice. We hope that through the Blueprint, Roadmap and exemplars that this project will be able to inform wider developments in the sector.

We hope to track this impact through emerging communities of practice. There is a commitment to frequent exchange of information with other institutions, including the other institutional JISC funded research data management projects, through a "Birds of a Feather" network. There is also a commitment regionally to explore technical issues and the potential of shared services with UCL and Oxford. All EPrints developments will be shared with the international EPrints community and take up of new applications will be monitored.

### **3 Conclusions**

There are a number of conclusions the project team have drawn from this work. These particularly relate to the extent to which it is possible to adopt either theoretical model to broad-based academic institution with a multi-disciplinary research profile.

#### *General Conclusion*

There is very considerable scope for a collaborative model in delivering effective research data management institutional level between senior managers, academics and researchers and service providers. The research lifecycle model is complex and requires a complex approach which can only be effectively delivered through partnership. This relies upon a higher level of training and awareness across the community so that the potentially very high demand on individuals to manage data effectively is eased through the development of knowledge and expertise. It is also true that the potential cost of fully effective data management both in terms of the resource required for storage curation and preservation, and the human resources for supporting and maintaining that infrastructure, is very considerable.

#### *Conclusions relevant to the wider community*

Much of the initial work on research data management has been focussed on very detailed analytical work describing elements of the research lifecycle and the structure of data. These have been hugely valuable in raising awareness of both the complexities and the potential vulnerability of research data. The application of some of these models is very challenging, and possibly beyond the scope of an individual institution. It is certainly beyond the scope of individual researchers.

There is therefore value in considering the role of shared services in the widest possible sense to add value. Shared services already include some national data centres, but these do not cover all disciplines. Shared services could also include regional consortia and on a technical plane, access to Cloud Computing. There are significant problems with these approaches not least authentication and control of data, and the robustness of some of the Cloud-based solutions which are currently available, but nonetheless as the amount of data expands exponentially, it is difficult to see that individual institutions, however well-funded, can sustain the cost. We found within our surveys that researchers believe that they should keep all data forever, and the recent EPSRC requirement is for data to be kept for ten years following any request for that data to be released. The feasibility of these requirements has to be balanced against a realistic appraisal of the relative cost. Institutions are still assessing options in this complex environment of competing calls on large scale investment in storage and infrastructure. It is likely that institutions will continue to pilot solutions using a range of providers in the medium term. Our experience of the SharePoint and EPrints pilots highlighted the importance of metadata that facilitates a portable and flexible approach over monolithic structures.

#### *Conclusions relevant to JISC*

There are two areas where JISC can help to make an impact in this area for which to present range of funding is of huge value to the wider community. Firstly in the area of collaborative business models on the lines outlined above. Secondly in the area of promoting cultural change among the research community, which would include collaborative approaches to service delivery on training and advocacy. Fundamentally it is the researchers' behaviour that we need to influence, not through a culture of compliance but through a culture of collaboration.

## 4 Recommendations

### *Recommendations for the wider community*

Significant engagement with senior staff in the institution is essential. The complex nature of data management means that developments need to link into high level strategic planning, business modelling and cross-service delivery. This has to be endorsed at the highest level.

The process of gathering information through audit and iterating data management policy is an important part of cultural engagement. It is not just about the evidence and results, although these are an important foundation for decision-making.

It would be useful to share examples of curriculum development, with incorporation of best practice training into modules and programmes. It is also timely to share approaches to integrate data management training into Graduate Schools, as many institutions are at a similar stage of reviewing support for PGRs post Roberts funding. The DCC could play a useful role in collecting and promoting exemplars.

### *Recommendations for JISC*

It will be useful to support and grow user communities for new software or the use of existing software to support data management and curation. We will actively seek to contribute to the ePrints community and the growing number of institutions testing the potential role of SharePoint in data management. Open source apps may be developed that will also interface with commercial products like Sharepoint and the early use of cloud services will test a range of service provider options. Clusters of linked support for shared services and shared expertise will be important.

## 5 Implications for the future

The data policy exemplar contributes to a growing number of data policies which are helping build momentum internationally for good data management practice. It provides an institutional approach to complement the principles and policies from Research Councils, Wellcome and other funders.

Once an institutional data repository and catalogue is rolled out we will continue to seek ways to maximise engagement and incentivise deposit. Further development of apps, some discipline specific, could encourage ingest and support the capture of more complex metadata. There also needs to be more work on the exchange of metadata and data between institutional and funder repositories or between institutional data publication repositories and archives.

A collective approach to training could be beneficial for those working to further develop and embed policy and guidance, offering services to support data management planning and other professional areas of expertise. This could usefully be linked to the "data manager" professional profiles currently being developed by the DCC to support the DataTrain JISC funded activity. This could be a mechanism for exploring the potential of new roles and investigate the challenge of supporting both discipline specialism and an integrated cross-service approach to skills. The aim is to expand the capacity and range of skills in all areas of expertise, from librarians to bid managers to technicians, to provide support that is robust as well as specialist. Nationally the enhanced skills sets should assist with the fluidity of best-fit movement across the sector.

Business-modelling has been a key component of this project and storage issues have been highlighted as a primary concern. As we move into an embedding phase with a significant roll-out of training activity we feel there is more work to be done on modelling the costs of embedding data management training into the curriculum and Graduate School activities. This human cost could be just as significant as storage and infrastructure, but requires additional investigation. This "big picture" approach to full institutional costs will support an opening out of data management as part of the overall information landscape, to supplant a discrete budget-line or ring fenced activity.