

JISC

BENEFITS FROM THE INFRASTRUCTURE PROJECTS IN THE JISC MANAGING RESEARCH DATA PROGRAMME

Final Report

Version 5.0, September 2011

Neil Beagrie

Charles Beagrie Ltd.

2 Helena Terrace

College Street

Salisbury SP13AN

Tel: 01722324925

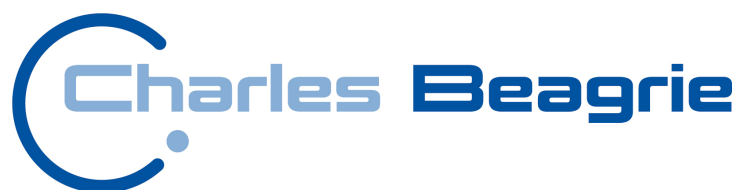


Table of Contents

1. Executive Summary	3
2. Introduction	7
2.1. Background	7
2.2. RDMI Benefits Case Studies	7
2.3. RDMI Project Business Cases	8
2.4. The RDMI Cost/Benefit and Business Case Support Project	8
2.5. The Integrated Data Management Planning Toolkit and Support Project (IDMP)	10
3. Overview of Projects' Benefits	11
3.1. Introduction	11
3.2. ADMIRAL	12
3.3. Institutional Data Management Blueprint	14
3.4. FISHNet	15
3.5. HALOGEN	16
<i>Direct Benefits</i>	17
<i>Indirect Benefits (Costs Avoided)</i>	17
<i>Possible metrics for these Benefits</i>	18
3.6. I2S2	18
3.7. Incremental	20
3.8. MaDAM	22
3.9. Sudamih	24
<i>Benefits from and potential metrics for training</i>	24
<i>Benefits from and metrics for providing 'Databases as a Service' (DaaS)</i>	25
4. The Support Programme Benefits	27
5. Sustainability and Business Cases	28
5.1. Introduction	28
5.2. Training and Guidance	28
5.3. Policy	30
5.4. Technical Infrastructure and Tools	31
<i>FISHNet</i>	31
<i>Sudamih DaaS</i>	32
6. Conclusions	34
6.1. Data Management Training and Guidance	34
6.2. Research Data Policy	34
6.3. Research Data Infrastructure	35
6.4. Outcomes	35
6.5. The Challenges in Defining Metrics for Research Data Management Projects	35
<i>Data Citation</i>	35
<i>Timescales and Attribution</i>	35
<i>Maturity and Critical Mass</i>	36
<i>Resources and Skills</i>	36
<i>Scope</i>	36
6.6. Assessing Benefits and Metrics in Future Programmes	36
7. Further Information	37
7.1. MRD Programme	37
7.2. RDMI Projects	37
7.3. MRD Programme Support Projects	37
7.4. Keeping Research Data Safe Reports and Tools	37
7.5. CARDIO and other Tools	38

1. EXECUTIVE SUMMARY

JISC's Managing Research Data programme has, with an investment of nearly £2M, funded a strand of eight Research Data Management Infrastructure (RDMI) projects to provide the UK Higher Education sector with examples of good research data management.

The RDMI projects have identified requirements to manage data created by researchers within an institution, or across a group of institutions, and then piloted research data management infrastructures at institutional, departmental or research group level, to address these requirements.

This report provides an analysis and synthesis of the benefits from this work identified by the eight RDMI projects in their benefits case studies, the benefits and enhancements that accrued to existing tools and methodologies from them, and the emerging business cases (as of April 2011) for sustainability being built by the RDMI projects.

The eight RDMI projects cover a wide-range of different contexts from a single university to multi-institutional disciplinary services. They also cover a range of infrastructure issues ranging from training and guidance, university research data policy, and technology issues such as tools, metadata, hardware and software.

For section 3 of this report, a shorter synopsis of the benefits and any potential metrics identified for them has been prepared for each of the projects from their benefits case studies. Section 4 provides a similar synopsis of the benefits gained via the two programme support projects. Tables are appended to this Executive Summary providing a summary listing of the benefits and metrics for them identified and described in greater detail by the projects in these sections.

To help assess future sustainability, projects have considered the costs, benefits and alignment with institutional strategy of any proposed solutions and provided examples of how universities may make effective, reasoned and costed decisions concerning the implementation of data management policies and infrastructure. It is not possible at this point to give a definitive view on costs and sustainability as this will accumulate and form over a longer time period. The RDMI business cases are starting to be finalised and reviewed by their institutions. In section 5, the emerging indications from 4 projects that had draft business cases by April 2011 are synthesised under three main themes for research data management infrastructure: training and guidance; policy; and technical infrastructure and tools.

In total, the projects have been able to accumulate an impressive body of evidence of the benefits (and potential metrics) from the programme. The projects were on average of 18 months duration so this report and the project work on which it is based, should be seen as a provisional assessment of those benefits. Their full impact may only be apparent and measured over a longer timespan.

Overall the cost/benefit support project emphasised the relative newness of impact and benefit assessment for research data management and corresponding challenges arising from this; the importance of providing templates and guidance on best practice to assist projects; and the implications in terms of resources and time for undertaking different levels of impact and cost/benefit analysis by projects. There is a need to further explore the metrics that can be applied to measure many benefits identified. These lessons should be valuable pointers for future programmes, institutions and their researchers.

University and research budgets are currently under considerable pressure to achieve more on lower funding. It is also anticipated that as a result of the Wakeham Review, the indirect costs element of research grants (which provide a significant part of the funding for research data infrastructure in universities) will be progressively reduced in coming years. The RDMI projects have undertaken a range of benefits case studies and analyses to demonstrate the varied ways in which research data management infrastructure can generate benefits for the HE sector: e.g. savings of time, more efficient research, better allocation of resources to active research, greater opportunities for sharing and reuse. Individually and as a group, it is hoped that these will contribute to an evidence base for improving data management practices and achieving further efficiencies in UK Higher Education and research.

Summary of Benefits Identified by the RDMI Projects

Benefits for Institutions

- New research funding and research opportunities
- Sustainability of research data infrastructure
- Economies of scale
- Better and larger publication output
- Integrated thinking around research data management
- Improved data management plans and policies
- Cost modelling to plan for increasing demand
- Prototyping technical solutions based on established technology
- Long-term road-maps for research data management
- Change to user practices
- Stimulating new networks and collaborations
- Mitigating organisational risks
- Increased awareness of funder requirements
- Sharing of exemplars, lessons learnt and best practice
- Impact and knowledge transfer

Benefits for Researchers/Research Teams

- Inspiration for new research
- Built-in support for research tasks
- Rapid access to results data and derived data
- Time and efficiency savings
- Increasing data management skills
- Increased awareness of relevant support, services and tools
- Enhancing finding and organizing of data
- Reliable citations to their data
- No loss of access to data as a result of Post Doc turnover
- Guidance and training for researchers embedded in Graduate Schools
- Supporting collective data collection/creation practices
- No re-creation of existing data
- Improved version control
- Secure back-up and reduced risk of data loss

Benefits for Research Support Services

- Better knowledge of the research data landscape

- Better awareness of researchers' needs
- Re-use of infrastructure in other projects
- Enhancements of and guidance for existing tools and development of new ones
- Improved data storage and data management
- Streamlined and updated advice and guidance
- Code developers able to enhance research software and tools
- Reduced risks of errors from manual transcription between systems
- Enhanced long-term stewardship
- Lower current and future preservation costs

Benefits for Scholarly Communication and Access

- Support for data publishing
- Enhanced data sharing and discovery and incentives to deposit
- Overcoming fears of data abuse
- Reducing effort to present data
- Greater consistency and standards between projects to enable data re-use
- Comprehensiveness of available information
- Improved quality of research by locating better, more relevant information
- Improved quality of research by linking relevant materials
- Re-purposing data for new audiences
- Re-purposing of methodologies
- Protecting returns on earlier investments
- Support of nationally important datasets

Summary of Metrics Identified by the RDMI Projects

On the whole, potential metrics are less fully discussed and developed in the project case studies than identification of benefits due to the challenges outlined and discussed further in section 6.5. However, despite the short timescale of the projects and other challenges, many were able to identify potential metrics that have or would provide measurement of specific benefits for different stakeholders including:

Benefits Metrics for Institutions

- New research grant income
- Number of research dataset publications generated
- Number of research papers
- Improvements over time in benchmark results (e.g. repeats of AIDA benchmarking or surveys of awareness of relevant support services or funder requirements)
- Cost savings/efficiencies for central services and/or departments
- Re-use of infrastructure in new projects

Benefits Metrics for Researchers/Research Teams

- Increase in grant income/success rates
- Increased visibility of research through data citation
- Percentage improvement in routine back-up of data
- Average reduction in waiting time (time latency) for data requests
- Average time saved in research data management and grant proposal activities
- Percentage improvement in range/effectiveness of research tool/software

Benefits Metrics for Research Support Services

- Percentage of potential user community that takes up services
- Number of data deposits with a repository
- Number of downloads of a dataset(s) within a repository
- Activity based costing methods (e.g. using KRDS activity model to benchmark activity based costs over time)
- Results of user feedback forms
- Number of times different researchers collectively create/maintain a dataset via the repository

Benefits Metrics for Scholarly Communication and Access

- Number of citations to datasets in research articles
- Number of citations to specific methods for research data management
- Number of datasets deposited with enhanced metadata
- Percentage increase in user communities
- Number of service level agreements for nationally important datasets

2. INTRODUCTION

2.1. BACKGROUND

The research process is enhanced by managing and sharing research data. Good research data management practice allows reliable verification of results and permits new and innovative research built on existing information. This is important if the full value of public investment in research is to be realised. These principles have been recognised by key stakeholders: most Research Councils now have policies in place which encourage or mandate the creation of a research data management plan and the deposit of research data in a recognised data centre where such exist. Many leading journals require underlying datasets also to be published or made accessible as part of the essential evidence base of a scholarly article.

As a result, higher education institutions are coming under increasing pressure to manage the research data generated by their researchers that cannot be curated by subject-based data centres – and many are unsure how to proceed given the absence of clear good practice. To address such concerns, JISC's Managing Research Data programme has, with an investment of nearly £2M, funded a strand of eight Managing Research Data Infrastructure (RDMI) projects to provide the UK Higher Education sector with examples of good research data management.¹

The RDMI projects have identified requirements to manage data created by researchers within an institution, or across a group of institutions, and then piloted research data management infrastructures at institutional, departmental or research group level, to address these requirements. In order better to understand the investment and change that may be required, cost-benefit analysis has been included in the projects' work.

University and research budgets are currently under considerable pressure to achieve more on lower funding. It is also anticipated that as a result of the Wakeham Review², the indirect costs element of research grants (which provide a significant part of the funding for research data infrastructure in universities) will be progressively reduced in coming years.

2.2. RDMI BENEFITS CASE STUDIES

Following review of previous benefits and assessment work undertaken as part of preparing the Programme Guide³, it was decided to follow the case study approach to assessing benefits in the RDMI projects. Recent work reviewing impact case studies for the pilot Research Excellence Framework exercise⁴ had suggested the use of templates and best practice guidance was needed for such work. These were therefore prepared to support the RDMI project in preparing their benefits case studies. This proved to be a successful approach for the programme.

The RDMI projects have undertaken a range of benefits case studies and analyses to demonstrate the varied ways in which research data management infrastructure can generate benefits for the HE sector: e.g. savings of time, more efficient research, better allocation of resources to active research, greater opportunities for sharing and reuse. Individually and as a group, it is hoped that these will contribute to a broader case for improving data management practices in UK Higher Education and research. Summaries of the case studies are provided in this report.

¹ Eight projects were funded from the JISC 07/09 Grant Funding Call to form the Research Data Management Infrastructure programme strand: <http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmi.aspx>

² <http://www.rcuk.ac.uk/documents/reviews/fec/FECReviewReport.pdf>

³ http://www.beagrie.com/DMIcost&benefit_programmeguidev1.pdf

⁴ <http://www.hefce.ac.uk/research/ref/impact/>

2.3. RDMI PROJECT BUSINESS CASES

A similar approach was taken to supporting the projects in the preparation of their business cases. Initial desk research was undertaken to identify and analyse exemplars of business case formats from the UK university sector and wider public sector as appropriate. Two generic business case templates were produced in Microsoft Word for preliminary and detailed business cases respectively with guidance in accompanying notes. Its primary audience was the JISC Research Data Management Infrastructure projects undertaking sustainability planning that did not have an institutional template to be followed for presenting an internal business case. However the generic guidance and specific links to other tools was also helpful to others, particularly if they were completing a business case for the first time. A synthesis of four existing business cases is provided in this report.

2.4. THE RDMI COST/BENEFIT AND BUSINESS CASE SUPPORT PROJECT

The RDMI projects have been assisted in this work by the cost/benefit and business case support project. Its role was to provide RDMI projects with methodologies such as the Keeping Research Data Safe (KRDS) activity-based cost model and KRDS Benefits Framework and assistance for analysis of the costs and benefits of their implementations. To do this the programme support project undertook the following tasks:

- Prepared a programme guide on cost/benefit analyses for research data⁵ covering current and recent work on cost/benefit and impact studies for research data and research data infrastructure;
- Updated the KRDS tools and provided new guidance on their use for the RDMI projects and the wider community⁶;
- Contributed to two programme general meetings and ran a dedicated workshop on costs/benefits with projects in November 2010;⁷
- Undertook initial site visits in conjunction with the Digital Curation Centre (DCC) and the Integrated Data Management Planning Toolkit and Support Project (IDMP) and then provided ongoing remote support and advice to projects as they progressed;
- Provided templates and best practice guidance to aid the projects in preparing their Benefits Case Studies and their Business Cases;
- Writing this report with a summation and analysis of cost/benefit work by the projects.

A major focus of the support project was use of the KRDS toolset and user guidance. The Keeping Research Data Safe projects' (KRDS1 and KRDS2) reports have been extremely well received by the community – KRDS1 alone was JISC's most downloaded publication in 2008. However the outcomes such as case studies and guidance were split over two long reports (169 and 88 pages respectively plus appendices and supplementary material). Feedback had suggested the need to prepare two syntheses of the report – a four page "Factsheet" and a succinct "KRDS User Guide" giving key implementation guidance and links to prepared extracts such as case studies from the reports. The KRDS User Guide was prepared for the programme and is an edited selection and synthesis of the KRDS reports combined with newly commissioned text and illustrations. The new KRDS User Guide had 875 downloads in the 4 months from publication and an accompanying KRDS Factsheet has 4,100 downloads by April 2011.

⁵ See http://www.beagrie.com/DMLcost&benefit_programmeguidev1.pdf

⁶ See <http://www.beagrie.com/krds.php>

⁷ See <http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmevents/mrdworkshop.aspx>

However, the KRDS Benefits Framework tool was used by a wide range of the MRD projects and revised in light of their feedback. The tool defines three dimensions and a framework to illuminate the broad outlines of the benefits that digital preservation investments for research data potentially generate. A short breakout session at the November cost/benefit workshop for projects from the JISC Managing Research Data Programme was used to review version 1 of the KRDS Benefits Framework. Feedback underlined the importance of the topic to attendees and suggested some additional guidance and revision of Dimension 3 (who benefits internal/external) of the Benefits Framework along lines of researcher and central services/institutional. This was incorporated into version 2.0 of the Framework.

The KRDS Benefits Framework		
Dimension 1 (Type of Outcome)	Direct Benefits	Indirect Benefits (e.g. costs avoided)
Dimension 2 (When)	Near-Term Benefits (up to 5 years)	Long-Term Benefits (5 years+)
Dimension 3 (Who Benefits)	Internal Benefits	External Benefits

Expansion of KRDS Benefits Framework Dimension 3 (Who Benefits) Sub-divided by a University's Stakeholders					
Internal Benefits			External Benefits		
Researcher	Research Group	Institution	Research Funder	Discipline	Others (e.g. NHS, etc)

2.5. THE INTEGRATED DATA MANAGEMENT PLANNING TOOLKIT AND SUPPORT PROJECT (IDMP)

Guidance to the MRD projects on requirements analysis was provided by the Integrated Data Management Planning Toolkit and Support Project (IDMP).⁸ The IDMP Toolkit and Support project sought to provide initial support to the MRD projects in their requirements gathering and benchmarking activity. A series of site visits was carried out by the IDMP team in early 2010 to provide the new projects with dedicated support in their application of the Assessing Institutional Digital Assets (AIDA), Data Asset Framework (DAF), and Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) toolkits.

The feedback the support project received on the applications of the individual tools has helped to shape the development of the Collaborative Assessment of Research Data Infrastructure and Objectives (CARDIO) tool⁹. CARDIO integrates the concepts of the aforementioned tools rather than their workflows which can be complex and time consuming to undertake. CARDIO draws together the broad concepts of risk, maturity and capacity at both data and infrastructural levels and serves as a jumping off point for those wishing to undertake more detailed DAF or DRAMBORA assessments. CARDIO will employ a rating system, however the emphasis will lie in building consensus amongst the various stakeholders. CARDIO aims to provide objectively meaningful benchmarks along with targeted, practical recommendations for improvement. These recommendations are drawn from the CARDIO knowledge base which makes visible the anonymised legacy data that were previously locked away within the individual toolkits.

⁸ <http://www.jisc.ac.uk/whatwedo/programmes/mrd/supportprojects/idmpsupport.aspx>

⁹ <http://cardio.dcc.ac.uk/>

3. OVERVIEW OF PROJECTS' BENEFITS

3.1. INTRODUCTION

Eight JISC MRD infrastructure projects completed project benefits case studies. Each benefits case study is approximately 5-7 pages in length and can be accessed in full at http://www.jisc.ac.uk/whatwedo/programmes/mrd/outputs/benefit_studies. For this report, a shorter synopsis of the benefits and any potential metrics identified for them has been prepared for each of the projects from their benefits case studies. The eight projects cover a wide-range of different contexts from a single university to multi-institutional disciplinary services. They also cover a range of infrastructure issues from training and guidance and university research data policy, to technology issues such as tools, metadata, hardware and software. A summary of the context and main foci for each project is provided in the table below.

Summary of the Context and Main Foci for each RDMI Project					
Project	Context	Training & Guidance	Policy	Technology	Page
ADMIRAL	University of Oxford Bio-Sciences/Life Sciences			✓	12
Blueprint (IDMB)	University of Southampton university-wide	✓	✓	✓	14
FISHNet	Freshwater Biological Assoc. Bio-Sciences			✓	15
HALOGEN	University of Leicester Humanities and Genetics		✓	✓	16
I2S2	Multi-institution/national facilities Physical Sciences			✓	18
Incremental	University of Glasgow/Cambridge university-wide	✓			20
MaDAM	University of Manchester Life Sciences/university-wide		✓	✓	22
Sudamih	University of Oxford Humanities	✓		✓	24

3.2. ADMIRAL

ADMIRAL (A Data Management Infrastructure for Research Across the Life Sciences) was a project to facilitate the capture of research data and their subsequent publication via an institutional repository (Oxford Databank). It was conducted by the Image Bioinformatics Research Group in the Zoology Department of Oxford University, the Oxford Bodleian Library Service, and the British Library. Its starting point was a shared file system that researchers can use immediately with little or no introduction, and which provides an immediate benefit of daily data backup.

From the project's initial interviews and data audits with researchers, Admiral found their current practice for research data management to be somewhat ad-hoc, with extreme variability between research groups and even between researchers within a single group. Research group leaders had some awareness of the need to manage and preserve data, but for other group members (post-docs, graduate students) this issue was very much secondary to their immediate research interests and their goal of publishing papers in high-impact journals.

Little thought was given to the eventual publication of their own research datasets. This was, in part, due to there being little recognition among researchers that research data are, in their own right, serious academic outputs, and also because there is a tendency among some researchers to regard their data as a personal resource for their own exploitation, rather than something to be shared.

This attitude is changing slowly - and the project staff have been attempting to catalyse this sea-change in attitudes through the ADMIRAL Project itself - as research councils and other funders publish data sharing policies and require that research data management plans are part of any submitted research proposal. For example, the BBSRC have published a data sharing policy¹⁰ that states:

- "BBSRC expects research data generated as a result of BBSRC support to be made available with as few restrictions as possible in a timely and responsible manner to the scientific community for subsequent research"
- "All applications seeking research grant funding from BBSRC must submit a statement on data sharing. This should include concise plans for data management and sharing as part of research grant proposal or provide explicit reasons why data sharing is not possible or appropriate."
- "the BBSRC 'Safeguarding Good Scientific Practice' document states that it expects primary data to be securely held for a period of ten years after completion of a research project". (See also: BBSRC Statement on Safeguarding Good Scientific Practice¹¹.)

The broad case for sharing research data, and the impediments to realizing the benefits of such sharing, were documented in some detail in a recent Nature article, Empty archives, by Bryn Nelson¹². While the small scale of the ADMIRAL Project was such that it could not hope to meet all of the indicated data sharing challenges on its own, it has addressed several of the issues described in the Empty archives article:

Finding and organizing data. ADMIRAL aims to make acquisition of data at source as easy as possible for the researchers. The system is designed to allow easy management, retrieval, annotation and addition to the data ('sheer curation' and 'curation by addition'). It will be possible to show progress on this front by having real research datasets, from small research groups routinely saved to and securely backed up in the local ADMIRAL file system, and by having selected datasets lodged in the Oxford Databank for long-term archiving and Web publication after an optional embargo period.

¹⁰ <http://www.bbsrc.ac.uk/organisation/policies/position/policy/data-sharing-policy.aspx>

¹¹ http://www.bbsrc.ac.uk/web/FILES/Policies/good_scientific_practice.pdf

¹² Data Sharing: Empty Archives, Bryn Nelson, Nature 461, 160-163 (2009), doi:10.1038/461160a:
<http://www.nature.com/news/2009/090909/full/461160a.html>

Overcoming fears of data abuse. Researchers are concerned to retain control of the data they produce and to manage easily who may have access to data before publication. It is also important to be able to manage whatever embargo periods may be allowed or appropriate. ADMIRAL aims to address this by allowing researchers to retain control over exactly what is published when. Each of the local ADMIRAL instances, one for each research group, is entirely private to that group.

Reducing effort required to present data. ADMIRAL supplies tools that provide useful views over the data and simultaneously automate acquisition of the information that makes possible such views. Progress will be demonstrated by having examples of additional metadata lodged with datasets that have been collected as the researchers analyse/organize their data to support journal publications.

Coordinated data storage. In many research groups or departments, the current practice of uncoordinated, unmanaged data storage, leads to effective data loss. 'All too many observations lie isolated and forgotten on personal hard drives and CDs, trapped by technical, legal and cultural barriers.' ADMIRAL does not directly address legal or cultural barriers, but progress on the technical barriers is being demonstrated through the use of common, shared storage mechanisms and systems. The ADMIRAL system makes data retrieval and use easier while ensuring secure backup.

Improving interpretation of shared data. For data to be shared and reused requires adequate contextual information. ADMIRAL does not address the issue of standards for primary data by imposing specific formats: indeed, they believe that researchers should be free to use formats that best serve their research needs. However, ADMIRAL does use standard formats and vocabularies for the descriptive metadata that accompany the datasets. The benefits of this will be demonstrated through the deposited datasets having associated metadata that allows a single query to retrieve information about multiple datasets.

Ability to cite shared data. There will be little recognition for good practice curation and sharing of research data unless the practice of citing data used is better encouraged and standard formats and mechanisms for data citation developed. ADMIRAL addresses this by working with the British Library to provide for Digital Object Identifiers (DOIs) to be assigned to published datasets. This will be demonstrated by having datasets published with assigned DOIs that can be cited within research papers and other datasets. More powerful evidence would be provided by examples of citations of such Oxford Databank datasets, but that is unlikely to happen within the time frame of the current JISC ADMIRAL project. However, it will, in principle, be possible to use future data citations to quantify the benefits of data publication through a facility such as the Databank.

In the near-term the key demonstrable outcome from the ADMIRAL Project will be a system that researchers are able and willing to use on a routine basis, which closes the loop between researchers' experimental data and data repositories.

Overlaying the file system with a Web access layer (HTTP and WebDAV) also allows Admiral to provide additional services that can be used to "improve" the data in small ways. The first such service is an easy-to-use service that allows selected datasets to be lodged with the Oxford Databank, the university's data repository service, from where the datasets can be published for wider, long term access. Additionally, the project's existing method of deploying separate independent ADMIRAL instances, configured using VMware as virtual machines within a common physical server and storage environment, bodes well for the future deployment of ADMIRAL systems and services more widely in other environments, including cloud-based environments, for the benefit of the broader research community. This work is now being taken forward in the DataFlow Project.¹³

¹³ <http://www.dataflow.ox.ac.uk/>

3.3. INSTITUTIONAL DATA MANAGEMENT BLUEPRINT

The Institutional Data Management Blueprint (IDMB) project aimed to produce a research data management framework across the whole institution of the University of Southampton. The project was structured around: determining current, and best, practice; analysing current capability; developing policy and governance; cost-benefit analysis; guidance and training for researchers; and developing short, medium and long-term goals for the institution. A key feature of this project was the engagement of stakeholders from across the University, including the Provost, Pro Vice-Chancellor for Research, Head of IT, University Librarian, and academics in the project team from the Schools of Archaeology, Chemistry, Electronics & Computer Science, and Engineering Sciences. Benefits from the project identified in the IDMB Benefits Case Study include:

Improved Data Storage and Data Management. An IDMB audit highlighted many of the day-to-day challenges faced by researchers across several disciplines. These range from simply not having enough storage, to wanting to catalogue their research data ready for uploading to national repositories. In order to tackle many of these issues, two pilot research data solutions have been implemented: a SharePoint 2010 system for archaeologists, and EPrints in a research data management role for the Southampton Nanofabrication Centre. These pilot implementations allow the university to see how more coherent, integrated, and intuitive data management solutions can be developed and deployed.

Streamlined and updated advice and guidance. To help researchers the Library and iSolutions (University central IT organisation) have updated the University web pages for research data management, creating clearer routes to those who can help, and designing a new, easy-to-use flowchart for researchers to help them with data management planning.

The IDMB project has been successful in bringing together key stakeholders from across the University to work together on a holistic approach to research data management. 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. Preliminary recommendations are put forward for short (one year), medium (one to three years), long (more than three years) term action. Key advances for the University include:

Better knowledge of the research data landscape: An audit has been carried out to find out how users manage their data, and how the University supports them. The AIDA (Assessing Institutional Digital Assets) toolkit has been used to benchmark current capability at the departmental/school and institutional level;

Integrated thinking around research data management: The project has developed 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;

Prototyping technical solutions based on established technology: 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. The project has undertaken 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;

Guidance and training for researchers, embedded in Graduate Schools: The project has addressed identified gaps in current provision such as: 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; there is little support for the creation of data management plans for research projects. As this requirement becomes more prevalent, researchers will need more support; finally 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. There is little central high-level guidance on best practices for data management;

Cost modelling to plan for increasing demand and requirements from funders: the project is modelling potential future costs and incorporating these into its business plan;

Long-term roadmap for research data management: the project has prepared a set of short-term, medium-term and long-term recommendations as part of a 10 year strategy for the university.

3.4. FISHNET

The Freshwater Information Sharing Network (FISHNet) aimed to understand the needs of freshwater scientists with regard to sharing data and to explore ways of meeting these needs. It was a joint project carried out by the Freshwater Biological Association (FBA) and the Centre for e-Research at King's College London. The first stage of the project involved gathering information from a variety of freshwater researchers, with the aim of finding out how they use and manage data and identifying their concerns over data protection, copyright and access control. This information has been used to implement a pilot data repository and web environment. Benefits from the project and possible metrics for them identified in the FISHNet Benefits Case Study include:

Enhanced data sharing and discovery. FISHNet improves data sharing by indexing all datasets submitted to the system and providing corresponding search facilities. This index also integrates information from the FBA's extensive freshwater library catalogue, image archive and hosted freshwater science websites, allowing users to discover not only datasets but also supporting material such as journal articles, books and images. The deposit of a dataset with the repository will fall into three categories, in which increasing levels of benefit are offered to the depositor in return for an increasing willingness to share the dataset. This is thus a "carrot" to encourage openness and data sharing. Measurements of the number of data requests or downloads from each category will provide feedback on the usefulness of this approach.

Easier access to data. FISHNet makes access to data easier by ensuring that the contact details of an appropriate person are always recorded for a given dataset and by providing a central point of access (the FISHNet repository). All downloads will be logged to provide a record of who has a copy and when it was taken. This will also provide a simple metric to determine whether the system is successful in the goal of increasing data sharing.

Addressing data sharing concerns. The FISHNet repository addresses the concerns about data sharing by making the decision to share or not to share it reside wholly with the uploader/owner, unless it has been freely released. This approach also allows for any licensing issues specific to that dataset to be addressed by making the request for and granting of access to a dataset into a bilateral transaction between two parties.

Comprehensiveness. Some institutions may not wish to have their data stored in an external system such as FISHNet as this may be against their existing policy. FISHNet provides for this by allowing metadata for a

dataset to be included in the repository, and indexed, without the dataset itself being present. Requests for such datasets can still be made through the FISHNet system and thus logged.

Improving data reusability and incentives to deposit. FISHNet addresses the issue of reusability by carrying out quality control on selected submissions. Repository staff work with the uploader to ensure that any data they submit conforms to appropriate standards and is readily interpretable to another user. This is likely to require effort on the part of the uploader and so a reward is needed to persuade them to make this effort. FISHNet will achieve this by providing a DOI, which will facilitate citation of the dataset and thus increase visibility of the depositor's work.

Supporting collective data collection/creation practices. FISHNet benefits a group of researchers who are collectively capturing data and wish to be able to contribute to a single, centrally-curated dataset in the FISHNet repository. One potential metric for this would be to record how often groups of users collectively create and maintain datasets.

Change to user practices. One potential benefit from FISHNet will be to change the practices of individual researchers as regards recording data by increasing standardisation and thus reusability. This is something that would be difficult to measure in practice, although it may be possible to define a benchmark for this behaviour by comparing the number of datasets deemed to be reusable (and thus receiving a DOI in the Orange category) upon first submission to FISHNet with those that require additional effort to change to a form deemed reusable. This would not be measurable for some time after the FISHNet project is completed however.

3.5. HALOGEN

The cross-disciplinary 'Roots of the British' collaboration between scholars in humanities and genetics at the University of Leicester (Wellcome Trust) seeks to interrogate the evidence for the migration and/or continuity of human populations in the British Isles in the distant past. The HALOGEN project (History, Archaeology, Linguistics, Onomastics and GENetics) supports the data management needs of the researchers involved and is thus establishing organisational best practice in terms of data management planning and the support of diverse cross-disciplinary research data.

The project was based on a pilot study focusing on the acquisition, cleaning, transformation and loading of three specific data sets into a single database. These were: the British Museum's Portable Antiquities Scheme; Nottingham University's Institute of Name Studies Key to English Place Names; and genetic Y chromosome data from the University of Leicester. It covered all phases of the project life cycle from requirements analyses through to the implementation of a practical, cost-effective database solution to meet the needs of the research team for data access, sharing, curation and preservation.

The database system was designed to exploit the existing research storage assets within the organisation in terms of hardware, software and technical expertise both within IT Services and other centres of excellence within the University, for example Library Services, Physics & Astronomy and Geography.

The development of central IT Services for the research community is a key part of Leicester's current IT Strategy. For many years central IT support at Leicester had been focused on supporting corporate services and systems. Through this project, the IT Services function has developed and enhanced the competencies needed to support designing and supporting multidisciplinary research databases and geospatial tools and techniques.

The Halogen project is perceived as extremely successful by the research community it supported. In that sense has helped IT Services promote itself within the organisation and to develop centralised IT support services targeted at the specific needs of the research community at Leicester.

The challenge of facilitating the increasingly common and wide ranging inter-disciplinary research projects (which may include external as well as institutional expertise) was addressed by: establishing a multi-disciplinary project team which represented the interests of all parties; identifying the need for and employing an experienced researcher with significant IT expertise (the Research Liaison Manager) to help facilitate interaction; and adopting pragmatic project management and development processes such as prototyping.

This approach was key to delivering a working database system to the research community.

The HALOGEN Pilot Project has been widely assessed as successful within the University of Leicester.

There is now a desire to build on this and make sure that the assets created by the project, both in terms of the physical database used by the researchers and, more generally, the model of engagement between the central IT service provider and researchers, are developed further.

In summary benefits and potential metrics identified by the project are as follows:

Direct Benefits

New research opportunities. Cross database work – bringing together and correlating new datasets alongside each other to seed new research outcomes. The pilot study acquired, cleaned, transformed and loaded 3 specific data sets into a single database.

Scholarly communication/access to data: Making two primary national resources (Key to English Place Names & Portable Antiquities Scheme) available for cross databasing with other research datasets such as the genetic Y chromosome data from the University of Leicester.

Verification, re-purposing and re-use of data. Cleaning & enhancing private or bespoke research datasets for reuse & correlation. Delivering increased transparency and excellent training for best practice in research data management.

Increasing research productivity. Building in cleaning, annotation and enhancement steps into normal research workflows so research datasets may be reusable immediately and interoperable with other datasets.

Stimulating new networks/collaborations by creating national visibility of the datasets and enabling data mining, correlation and new research sample selections.

Impact & Knowledge Transfer. The HALOGEN research IT infrastructure is being reused in other industrial multi-disciplinary projects e.g. the EU FP7 Mintweld (industrial engineering) & BRICCS National Health Service/University Trust data sharing projects.

Increasing skills base of researchers/students/staff. HALOGEN is supporting research data management training, research tool discovery and ArcGIS geo-spatial data visualization.

Indirect Benefits (Costs Avoided)

No re-creation of data. Researchers avoid valuable time needed to transcribe external data sources for reuse with their own research datasets.

No loss of future research opportunities. The HALOGEN Inter-disciplinary research platform is available for reuse as a service in other projects.

Lower future preservation costs: There are central Service Level Agreements in place to maintain the HALOGEN infrastructure and the input datasets. Preservation is not dependent on individuals alone and economies of scale are realised.

Re-purposing data for new audiences: Both internal & national research resources can become nationally available & reusable by new audiences, increasing exposure and impact.

Re-purposing methodologies: e.g. Geneticists learn better spatial correlation analysis techniques from GIS & astronomical e-research expertise.

Protecting returns on earlier investments: Nationally important datasets produced through research funded by Wellcome Trust, Leverhulme Trust, Arts and Humanities Research Council, the British Museum and the Universities of Leicester, Nottingham and University College London all now have Service Level Agreements to preserve and enhance these valuable research resources.

Possible metrics for these Benefits

New research grant income where a project seeks to use or reuse Halogen database infrastructure. The Roots of the British collaboration won £1.3m Leverhulme Trust funding for an interdisciplinary research programme over 5 years on the basis of reuse of the Halogen database.

Measurement of citations to Halogen or derived research datasets hosted on it in academic research papers.

Research productivity gains as measured by numbers of research papers generated and grant income per collaborator before and after participation.

Measures of reuse of infrastructure in other projects via IT Services and savings in individual departmental spend on research data storage and management. Numbers and volumes of data assets recorded in each case.

Support of nationally important datasets through Service Level Agreements with an institution rather than each individual originator.

3.6. I2S2

The Infrastructure for Integration in Structural Sciences (I2S2) project identified requirements for a data-driven research infrastructure in 'Structural Science', focusing primarily on the domains of Chemistry and Crystallography and involving researchers and service staff at multiple universities and national/international central research facilities. Two research data management pilots examined the business processes of research in structural science and addressed traversing administrative boundaries between institutions to national facilities in addition to issues of scale (local laboratory to national facilities, DIAMOND synchrotron and ISIS respectively). The primary or major benefits of implementing I2S2 identified by its two benefits case studies are:

Enhanced data management and long-term stewardship. The immediate beneficiaries are the core research teams and their staff and close collaborators. The changes that take place as a result of the project will immediately impact on their working practices and the benefits to their research that follow (better science, higher productivity) will be felt quickly;

Rapid access to results data and derived data. There is a substantial anticipated reduction in the latency of information access for derived data or results data. At the present time, the way to obtain such data from one's colleagues is to ask, and typically the latency cost is of the order of one day to receive the

data, which is borne by both the user and his/her colleague. It is estimated that implementation of I2S2 can reduce a typical one-day latency of data access down to an average of around five minutes: this is seen by the researchers as one of the key benefits of I2S2. It has very high impact on the efficiency and effectiveness of the research by shortening time gaps and leading to a more rapid progression towards final publication;

Increased productivity through time savings and increased efficiency. These are primarily appreciated by and visible at, the level of national facilities and services (or whole institutions) as economies of scale accumulate any time savings across multiple researchers, experiments and samples. The same benefits may be viewed as less significant or have lower impact at the level of individual researchers;

Better and larger publication output. The higher-education institutes, facilities and researchers will have a consequential benefit that accrues from a better and larger publication output;

Training. New users will benefit enormously by having ready access to a wide range of well-documented data examples for tutorials and practice studies. The wider user community is currently relatively small in I2S2 benefits case 2. However, it is anticipated to grow extremely rapidly in the UK, in part promoted through the availability of new instruments at ISIS and Diamond. This leads to important requirements for being able to scale training by reducing direct reliance on learning from a few existing researchers and increasing opportunities for self-learning for which availability of well-documented data is a key requirement;

Software and tool development. Code developers for the software and tools will benefit from having access to a wide and diverse range of well-documented data. The range and effectiveness of the analytical tools for the science can be enhanced if high-quality test data is available. There is a wide range of different use cases, and developers need access to a wide range of examples for testing purposes. Moreover, the number of use cases increases with time, and developers need to have access to an expanding range of examples and accompanying data to generate new versions to meet changing requirements;

Wider access and use. Facilities will benefit by providing access to results and derived data as part of their services. There will be easier retrieval or revisiting of experiments long into the future. Other research teams will benefit from having access to this data for new analysis or comparative studies;

Reducing risk. There will be less likelihood of errors in the safety or conduct of experiments as a result of better electronic information transfer and less manual transcription between systems;

Data publishing. The ability of data to be fully validated and therefore openly published without further context (i.e. journal article) and an increased visibility of data with a secured longevity will mean increased citation and greater long-term effectiveness of the research;

Knowledge transfer. There are companies that are now marketing lab-based x-ray sources optimised for obtaining PDF data. Researchers in Benefits Case Study 2 are collaborating with one, and for this company the benefits from I2S2 will be similar to that for researchers plus the ability to make demonstration data easily available. This is not merely good for one company's advertising; availability of lab-based equipment meets a real community need.

Potential metrics identified for these benefits are:

Service productivity and efficiencies. I2S2 has developed an activity model of the scientific research data lifecycle and associated tasks¹⁴. Using this to structure analysis, the National Crystallography Service activities that are expected to be significantly changed and impacted by I2S2 are being benchmarked to allow "before" and "after" time measurements. It should therefore be possible to calculate any work

¹⁴ <http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2-ResearchActivityLifecycleModel-110407.pdf>

efficiencies and time savings after full implementation of I2S2. Metrics for these benefits are particularly difficult to capture within the timeframe of short projects or the limitations of pilot implementations but benchmarks for longer-term evaluation have been established.

Extending, training, and self-starting the user community. In Benefits Case Study 2, the wider user community is currently relatively small. However, it is anticipated to grow extremely rapidly in the UK, in part promoted through the availability of new instruments at ISIS and Diamond. This leads to important requirements for training and ongoing code development, for which availability of well-documented data is a key requirement. The number of users and completed studies can be counted through the number of publications that cite the main program publication. A clear metric of success here will be an increase in the citation rate of the paper.

Higher work throughput and outputs through reduced latency in access to derived data and results data. The beneficiaries that will provide the benchmark here are the research teams and staff. The indicator of success is that they can turn an estimated typical one-day latency of data access down to five minutes.

Improved software and tools. The two markers of success are a) as new functionality is developed a suite of test data can be produced easily; b) that as new types of systems demand new functionality, there is the ability to add new data sets to the test suite. The number of use cases increases with time and the developers need access to an expanding range of examples and well-documented accompanying data to generate new versions of the software and tools to meet changing requirements.

3.7. INCREMENTAL

The Incremental project was a collaboration between Cambridge University Library and the Humanities Advances Technology and Information Institute (HATII) at University of Glasgow, with the primary goal of expanding research data management capacity in the Universities.

The project spoke to researchers, PhD students and computing officers across a range of disciplines and found common challenges throughout the research data lifecycle, finding many researchers:

- organise their data in an ad hoc fashion, posing difficulties with retrieval and reuse;
- store their data on all kinds of media without always considering back-up and security;
- are positive about data sharing in principle, but more reluctant in practice;
- believe back-up is equivalent to preservation.

These specific challenges represent a general lack of awareness of data management best practices (Challenge 1), causing substantial inefficiencies and risks in research (Challenge 2) (e.g. wasted time or serious data loss). Whilst there is some useful guidance within the Universities and on external websites that can assist with these challenges, researchers said it was typically hard to find guidance resources and the resources themselves were often too long or theoretical.¹⁵ There was also a lack of awareness of key institutional services (Challenge 3) (e.g. institutional repository, and support for remote access). Subsequent observations of researchers using our university websites gave further detail on how the format of resources (i.e. policies or long texts) could negatively affect usability.

¹⁵For a more comprehensive description of findings see the Incremental *Scoping Study Report* at: www.lib.cam.ac.uk/preservation/incremental/documents/Incremental_Scoping_Report_170910.pdf

In order to address these challenges, the project has provided clear and accessible guidance by creating centralised university data management websites, repurposing or reproducing existing resources, and signposting relevant services. The project has also created original resources to address gaps in provision, providing training courses and teaching resources for the University communities (and beyond), and making all of these resources easy to reuse and repurpose. Finally, project staff have reached out to important University offices, building connections, support networks, and relationships.

Because of their involvement in this project, the University Library and the institutional repository have gained substantial awareness of researchers' needs and concerns, and have a clearer picture of how to assist and engage with researchers. This understanding goes both ways; researchers and local support staff participating in the scoping stage of the project have gained a clearer understanding of how the Library and repository can help them. Overall the beneficiaries of the project can be classified as follows:

- Direct beneficiaries: researchers, staff who support researchers, university-level services.
- Indirect beneficiaries: researchers' colleagues and contacts.

The timescale for realising benefits is seen as:

- Near-term: immediate gains for participants in the scoping study and event attendees
- Medium-term: six months to two years for additional gains from spread of awareness.

Current and future anticipated benefits identified in the Incremental Benefits Case Study are summarised below:

Increasing research efficiency and mitigating risk. Incremental's initial scoping work found that most researchers are unaware of best practices for managing research data, some of which, they widely acknowledge, would increase research efficiency (e.g. consistent file naming when working with partners) and reduce risks (e.g. robust back-up procedures). Scoping study participants across the board told project staff that whilst they are interested in best practice guidance, they are not inclined to consult existing guidance resources, which are often difficult to find and tend to be long, theoretical and unengaging. Incremental has acted on concerns over the main University websites; for example, that pages are too long and text-heavy, and important information is often buried within large documents. There was a positive response to the new briefer guidance, which indicates Incremental's use of FAQ formats, short fact sheets, and visual aids such as flow diagrams will assist staff to locate relevant information and understand requirements.

Whilst the project has identified clear benefits in providing a centralised source of information and signposting for existing resources, it anticipates additional future benefits through encouraging personnel within University departments and external institutions to incorporate the resources into local websites and training. Accordingly, the project is publishing all Incremental resources and pages under Creative Commons licences and creating modular resources, including editable PowerPoint slides. Cambridge's DataTrain project (Archaeology and Social Anthropology) is already working to incorporate these resources into departmental training materials.

Increased awareness of relevant support and services and requirements. The scoping work was an opportunity to tell researchers about relevant University services of which they were unaware. Based on Incremental's findings from this work, project staff have provided details of relevant support in a prominent place on the web pages, making sure to provide contact information for tailored support. Incremental has also involved service staff in training and events to link researchers with support. Stakeholder evaluations of Incremental web resources confirm the project's findings from the scoping study, that researchers want centralised links to services, and they indicate that researchers will use Incremental's resources to gain

information on data management services as well as practices. Some also said that they will promote these resources locally. As information transfer often happens through peers (a belief confirmed by our scoping study) the project expects a wider impact via the researchers that have been engaged in its work. This is certainly indicated by the training feedback. In addition, Incremental anticipate that increased awareness of and capacity to meet funder and university requirements will reduce risk, increase efficiency, and strengthen relationships between researchers and their funders.

Increased awareness of researchers' needs. Incremental has also seen direct benefits from the process of undertaking the scoping and implementation phase of the project. The scoping work brought University support staff out into the research community. This increased the Incremental staff and their institutions' awareness of researcher practices and needs, and increased researcher and local support staff awareness of issues and support for managing research data. Support staff participation in Incremental seminars and workshops further strengthens this effect, and strengthens relationships between support services within the Universities.

Feedback has been positive about the web resources Incremental has developed at Cambridge and Glasgow, additional planned resources, and the events it has run to date. The centralised web pages have already had some practical application in supporting bid writing at Glasgow, and the Research Office at the University of Cambridge has agreed to point researchers toward relevant support services (e.g. the institutional repository), and project-made data management planning resources at the grant application stage.

3.8. MADAM

The MaDAM project at University of Manchester aimed to develop a pilot infrastructure for the better management of data across the research lifecycle, from data capture to storing, preservation and dissemination.

The MaDAM pilot user research groups come from two different biomedical domains at University of Manchester: 1) The Life Sciences Electron and Standard Microscopy group includes four sub-groups (overall consisting of 8 active core users plus some occasional users) who all work with large quantities of imaging data in diverse formats and resolutions. Within their specific research they use different methodologies and instruments (e.g. Standard, Cryo-Electron and 3D Tomography Electron Microscopes). 2) The research of the Medical Sciences Magnetic Resonance Imaging (MRI) Neuropsychiatry Unit (5 users) involves primarily brain imaging data from a number of distributed MRI scanners run by University, Wellcome Trust and NHS. This includes textual psycho-social data linked with MRI scans. The work with the pilot user groups is further complemented by information and requirements gathered from additional researchers and PIs within the domain, IT and experimental officers as well as research and data policy managers.

The basic immediate challenges for both user domains lie in storing their image data securely, backing-up, and providing researchers with an infrastructure to support their day-to-day data management. In the research lifecycle this pertains to the point after they have collected their image data and metadata from instruments. The instruments are 'firewalled' to insulate them from external networks for security reasons that mean the researcher must transfer the data using a portable device (e.g. USB and optical media) to their own PC.

At this stage the whole data set becomes entirely the researcher's responsibility and, in the absence of practical guidance around good management of data, every researcher has their own processes for back-ups, file management, annotation, metadata capture and storage locations and media for the short, medium

and long term. Raw data are manipulated and analysed through a series of steps, using a variety of computational and other techniques and software to produce various interim versions of processed and analysed data up to the point of creating outputs for publication and other forms of dissemination such as website material for public engagement.

There is currently no University of Manchester strategy specifically pertaining to research data management although there are policies from external (funding) bodies, and relevant internal policies around ethics, information security and data protection as separate themes. This lack of a supporting framework for policy at University level is likely due in part to the differences in needs, culture and politics of the different faculties and disciplines which operate almost as distinct entities and which may be difficult to reconcile. For the MaDAM pilot research groups this means their work practice regarding data management procedures or plans are quite diverse: mostly it is down to the single researcher, sometimes to the PI to set at least a minimum of standards.

The MaDAM pilot infrastructure includes a technical (hardware and software) system and curation and governance policies for research data management. Benefits identified in the MaDAM Benefits Case Study include:

Secure storage and back-up. Based on qualitative evidence of users' research practice mitigating the risk of losing data by providing a trusted, secure and central storage location with automatic back-up is a key feature of the MaDAM infrastructure – it is even more crucial for managing confidentiality and other ethical issues related to human data in the Medical domain. Further benefits lie in freeing up researchers' PCs and local storage.

Improved data management, search and annotation. The basic features make data and metadata highly visible and searchable in users' day-to-day research, thus making it easier to find and flag high quality data; and conversely helping to weed out non-useful files from the ever-larger amount of data. A list of thumbnails of an image set for example, automatically created by the MaDAM system, is usually the most suitable means for researchers to identify the relevant single/series of images from an experiment or scan. Linking data also helps to reduce redundancy from duplication and makes cross-studies more feasible. This all not only makes finding and cross-referencing data easier or in cases simply possible, it is also a huge benefit in terms of saving valuable time.

Data sharing. Although sharing data or open science have not been flagged as main requirements by the MaDAM pilot users at the moment, the MaDAM infrastructure is facilitating easier, more secure owner controlled data sharing. The integrated eScholar repository will give users seamless access to disseminate their research outputs. eScholar will also be the curation and preservation end point for research data in MaDAM.

Facilitating data re-use. Besides generally maintaining media and format accessibility for long term reuse, some data sets of value may not be recognised as such because of the time investment required to develop metadata, and these are at risk of loss through neglect because they are not currently flagged as potentially valuable. The project staff anticipate that MaDAM will play a role in making such data more visible, classifiable and re-discoverable as a mid- to long-term benefit.

Improving data management plans, policies and institutional settings. The MaDAM project has produced a Landscape Review document on Policies, Legal & Ethical Perspectives, Stakeholders and Institutional Settings. A Data Management Planning component has been included in the infrastructure to provide adequate metadata and guidance for its users. Users have been provided with assistance in the development of Data Management Plans (DMPs), both through a series of workshops, and also with in-tool support (eDMP). The in-tool support consists of a series of help pages which advise the researcher on data

management planning, funding body policies, contacts etc, as well as intelligent functions such as automatic review dates. MaDAM has also developed an interface with the Research Office's Research Management Systems to make the completion of DMPs less laborious, allowing us to populate part of the eDMP with information from such systems.

Sustainability and university strategy. There is a proposal for a wider Research Data Management Service (RDMS) at the University of Manchester, with the aim to roll out this service incrementally; adding research groups sequentially – starting with MaDAM as a demonstrator and with its findings being fed into the wider proposal. This proposal within a wider University research data management strategy is currently being explored and could open a sustainability route for MaDAM. MaDAM will also produce a cost-benefit case including IT Services' costing for initial (7TB) and longer-term (~32.5 TB) storage and hosting within the University of Manchester IT infrastructure. It has also been evident that in the end a cultural change is also needed for the proper support of domain specific data management plans, research practices, and research management policies in general, and this, inevitably, will take time.

Measuring benefits therefore might be best undertaken over a longer timescale and might include metrics for uptake of new tools and services; and deposit, citation and re-use of datasets.

Overall further evaluation and documentation of evolving and emerging patterns and behaviour of actual research practice in this context and hence of the uptake of the MaDAM infrastructure will be instructive beyond the project's initial life time. The MaDAM pilot work involving user feedback and prototyping will continue as part of the ongoing assessment for the further development of a data management and digital curation strategy for the wider proposal for a Research Data Management Service at the University of Manchester.

3.9. SUDAMIH

The Supporting Data Management Infrastructure for the Humanities Project (Sudamih) worked on two particular aspects of research data management infrastructure: training for researchers; and the development of simple and intuitive 'Databases as a Service' (DaaS) software which researchers can use to create, edit, and share their own research databases. The project worked with researchers from the Humanities Division at the University of Oxford to meet the particular needs of humanities scholars in the first instance. The University intends ultimately to expand the infrastructure developed to the other academic disciplines as well. The following benefits were identified in the Sudamih Benefits Case Study and derive from training, DaaS, or combination together of their effects:

Benefits from and potential metrics for training

The project identified eight benefits arising from the continued maintenance and dissemination of the training and learning materials developed by the Sudamih Project:

1. Time saved by researchers by locating and retrieving relevant research notes and information more rapidly
2. Improved quality of research by locating better, more relevant research information than would otherwise be the case
3. Improved quality of research by linking materials in such a way as to highlight connections and trigger new ideas
4. Improved comprehensibility of research information and data after long time periods, facilitating reuse

5. Better awareness and use of software tools to assist research management
6. Better awareness and uptake of central infrastructure services intended to help researchers, including technical help and assistance with funding bids
7. Reduced risk of data loss
8. Improved version control

Assessing the long-term impact of information management training is difficult, as whilst the benefits of improved organisational systems and techniques are likely to accrue over time, there are so many factors that could influence information management practices besides a specific training course or website, the benefits derived from any given set of training materials are hard to isolate and measure. Sudamih has therefore attempted to gather what information it can about more immediate short-term impacts, on the principle that if no short-term impact can be detected the likelihood of significant long-term benefits is also doubtful. Thus, whilst these benefits are not straightforward to quantify, the project has surveyed responses to the new training materials where possible to assess the level of impact they would need to justify the costs of continued provision.

The face-to-face courses run by the Sudamih project were extremely popular and effective, at least insofar as effectiveness can be measured by altering behaviour. Asked 'have you/will you change any aspects of your own information management as a result of the course?', 23% of respondents said that they had made or would make 'significant' changes, 69% said that they had or would change at least one or two aspects of their current practice, and 8% that they were considering changes. None of the respondents reported that they were not even considering changes after attending the courses.

In an attempt to measure benefit 1 (time saved by researchers by locating and retrieving relevant research notes and information more rapidly) Sudamih asked course attendees to estimate how much of their time spent writing up their research outputs is actually spent looking for notes/files/data that they know they already have and wish to refer to. The average was 18%, although in some instances it was substantially more, especially amongst those who had already spent many years engaged in research (and presumably therefore had more material to sift through). This would indicate that there is at least considerable scope to save time (and improve research efficiency) by offering training that over the long term could improve information management practices.

Sudamih undertook a similar approach to address benefits 6, 7 and 8 – the related benefits aimed at reducing the risk of data loss and improving version control. 61% of those course attendees who completed the feedback survey indicated that they had at least on occasions lost information or data that they had wished to refer back to. Most respondents did not attempt to quantify precisely how many days' work they had lost, and none had suffered any catastrophic loss, but two respondents estimated that they had lost at least a month's work during their five or six years of research. By reminding researchers about, or introducing them to, centrally provided back-up and security services and tools and methods for file synchronization and version control, it can reasonably be expected that these figures can be reduced.

Benefits from and metrics for providing 'Databases as a Service' (DaaS)

The anticipated benefits identified as arising from the use of the DaaS are as follows:

1. Improved sharing and reuse of data
2. Improved data security (storage; access & identity management)
3. Less duplication of data
4. Faster preparation of structured data

5. Greater technical support efficiency
6. Economies of scale due to centralized hosting
7. Greater data consistency between projects facilitating the repurposing of data and mash-ups
8. Data becomes more reliably citable due to use of DOIs
9. Strengthened research grant applications
10. Greater awareness of existing data provides inspiration for new research

As with the training outputs of the Sudamih Project, it is not possible to quantify many of the benefits arising from the uptake of the DaaS within the timescale of the project itself. Given that the DaaS is only just taking shape in its pilot form as the project draws to its conclusion, the Sudamih project cannot yet fairly compare its performance against more established database management software used by researchers. The project has, however, tried to measure future demand, and we have also asked humanities researchers who work with structured data to give us their assessment of the intended service.

Feedback from Humanities Reserachers: The project has evidence that there is considerable demand for the capabilities offered by the DaaS thanks to feedback from a workshop organised by Sudamih entitled 'Databases in the Humanities – Where Next?' The workshop was used as an opportunity to explain the features of the DaaS. Attendees (who were mostly researchers in the humanities who already had an interest in research data and databases) were asked whether they would consider using the DaaS (or an equivalent) once it was fully operational to develop or host a database. 56% of respondents indicated that they would seriously consider it, whilst a further 44% said it was a possibility. Respondents gave various reasons why they thought the DaaS might interest them, including: expanding upon projects begun using personal database software such as Access; sharing data input; for hosting small-scale personal projects; and as a cheap long-term host for projects that are no longer otherwise supported.

Assessing potential savings. Whilst the university has not been able to measure all of these benefits within the timescale of the Sudamih Project, it has assessed potential savings deriving from the use of the DaaS by gathering information relating to the costs involved in creating, sharing, and maintaining a large research database. It did this by using the Roman Economy Project as a case study. The Roman Economy Project is currently working to combine a number of databases into a coherent whole with a public search interface and is testing the DaaS as a means for so doing. The university measured costs of developments so far, including data gathering, reorganisation, and current hosting models and contrast these with the anticipated costs of supporting the DaaS.

Relevant comparators. Sudamih was able to get a clearer assessment of some of the other potential benefits (such as improved support efficiency and economies of scale) by talking to support teams in the Computing Services.

Details of this work are presented below in **Section 5.4 Technical Infrastructure and Tools, Sudamih DaaS.**

Whilst Sudamih has already gone some way to illustrating need and demand for its training and DaaS outputs, more work remains to be done to verify the anticipated benefits and, where possible, to quantify them. It will not be possible in many instances to produce estimates of benefits in financial terms, but they should be able to give clear qualitative justifications for the proposed services.

4. THE SUPPORT PROGRAMME BENEFITS

The MRD Projects were supported in their requirements analysis work by the Digital Curation Centre (DCC) and in their cost-benefits analysis and preparation of benefits case studies and business cases by Charles Beagrie Ltd. These are briefly described in the Introduction (see sections 2.4 and 2.5).

The Assessing Institutional Digital Assets (AIDA) self-assessment toolkit, which began life as an assessment tool applying to *all* digital assets, was recast in 2010 to be used specifically for the analysis of management of research data in the context of the MRD programme. It has had some take-up from three projects: Blueprint and Sudamih (both RDMI projects) and Gravitational Waves (part of the wider MRD programme). Its "recast" version has, with some refinement, formed the backbone of the new Collaborative Assessment of Research Data Infrastructure and Objectives (CARDIO) tool developed in the course of the IDMP project which will help institutions to assess and benchmark their data management service provision and maturity.

This has been a two-way interaction with benefits also accruing to the support projects and their tools. The findings which have emerged from the Projects' requirements analyses and their use of data management planning and benefit analysis tools has fed into work to enhance and integrate them for the wider community. Feedback and benefits gained from working with the RDMI projects and from enhancements to the toolsets are detailed by the two support projects below. In summary benefits identified are:

A new Collaborative Assessment of Research Data Infrastructure and Objectives (CARDIO) tool has been developed that will help institutions assess and benchmark their data management service provision and maturity;

Refining the scope and most effective application of the tools. The support projects learnt from the applications of existing tools for requirements analysis such as AIDA, DAF and DRAMBORA that starting with a small, well-defined scope is more successful than starting on a broader scale;

The value-added by the assessment process itself. In most cases the wide range of stakeholders involved in these assessments (researchers, central support, IS) had little contact with each other on a day to day basis and didn't always understand each other's concerns and requirements. Communications between these groups often became difficult due to a lack of shared understanding of data management and curation terms and issues;

Improved versions and guidance for the Keeping Research Data Safe (KRDS) Cost/Benefit Tools. As part of the programme, a new KRDS User Guide was produced providing a synopsis of the two KRDS reports. This had 875 downloads in the 4 months from publication and an accompanying KRDS Factsheet 4,100 downloads by April 2011;

Assessing costs by activity-based costing is too difficult for most projects to do themselves although they can learn from findings of others who have;

How to assess benefits. Assessing benefits and impact from research data management is still novel and often very challenging for projects particularly if they are of relatively short duration. However a case study approach supported by templates and guidance on best practice proved valuable and was successfully applied by all 8 RDMI projects.

5. SUSTAINABILITY AND BUSINESS CASES

5.1. INTRODUCTION

JISC required RDMI projects to report on findings throughout their projects, to include a consideration of the benefits of effective data management provided by the solutions implemented, and a business case to assess and support their sustainability. To help assess future sustainability, projects would consider the costs, benefits and alignment with institutional strategy of any proposed solution and provide examples of how universities may make effective, reasoned and costed decisions concerning the implementation of data management policies and infrastructure.

As noted in the previous chapter, the cost/benefit support project prepared two generic business case templates and guidance to support projects in pursuing funding from their institution to sustain their outputs following completion of their JISC grant. The reason that two templates were compiled is that in many universities/research institutes the principles of PRINCE2 Project Management Methodology are used. If applied, the business case can become a two-stage process with the first being a summary or proposal template to obtain management commitment and approval for the second stage, a more detailed and further developed business case. For some developments the benefits of a more “agile” approach is increasingly being recognised. This normally relates to a method of project management particularly relating to software development where it is necessary to have a highly responsive approach to change but it is more than likely that it will still need to go through the normal process of business case development, submission and approval.

Compilation of business cases for several RDMI projects, and the process of institutional review and revisions for all of them, is therefore still ongoing and not complete.

It is not possible at this point to give a definitive view on costs and sustainability as this will accumulate and form over a longer time period. However the RDMI business cases are starting to be finalised and reviewed by their institutions and provide initial indications of costs, benefits and alignment with institutional strategy of proposed solutions. The emerging indications from 4 projects (FISHNet, MaDAM, Incremental, and Sudamih) that had draft business cases by April 2011, are synthesised below under three main themes for research data management infrastructure: training and guidance (Sudamih and Incremental); policy (MaDAM); and technical infrastructure and tools (FISHNet).

5.2. TRAINING AND GUIDANCE

Given the importance of managing one’s research data properly and the potential costs of failing to do so, the Sudamih and Incremental projects’ work on research data management training turned out to be unexpectedly pioneering. For example, whereas at the start of the project Sudamih had envisaged creating a suite of training materials largely by recombining and customizing existing training materials from elsewhere, it became apparent very early on that there was very little currently available in an appropriate form. It therefore needed to produce much of the content from scratch.

The Sudamih Project ended on 31 March 2011, having produced and piloted four significant sets of training/learning outputs: two courses with accompanying course-books, slides, and step-by-step practical exercises, intended for face-to-face delivery; a suite of interlinked content written for the IT Learning Programme’s ‘Research Skills Toolkit’ website at the University of Oxford; three slide packs for use in researcher induction sessions; and a data management factsheet designed to accompany the Humanities Division’s ‘Managing the D.Phil.’ course.

The business case for Sudamih training points out that its training outputs were well received by researchers and project stakeholders. In FEC terms approximately £40,000 has been invested in developing and trialling the materials. They had been developed specifically for use in existing contexts within the University's training infrastructure, although a 'non-Oxford' version of each of the outputs has also been created for use by the broader HE community (as mandated by the JISC).

Eight benefits have been identified as arising from the continued maintenance and dissemination of the materials developed (see section 3.9). Whilst it is not easy to quantify all of these benefits financially, Sudamih have estimated where possible the impact that the training would need to have in order to cover the costs of continuing to provide the training.

The total future cost to the IT Learning Programme at Oxford University Computing Services of maintaining and delivering the training and learning materials comes to £4,080 per annum. Results of the cost/benefit analysis of reducing information discovery and retrieval times by even 2% and reducing data losses (estimated at 1.15% per annum from information gathered from participants) by half, suggest that these outcomes alone would quickly and substantially repay the training costs incurred.

The Incremental project ended on 26 April 2011. Both project partners (the Universities of Cambridge and Glasgow) have made internal business cases to their institutions. The results and lessons from these are documented in a separate report for the benefit of the JISC Managing Research Data programme¹⁶.

Business cases have been made to Library and DSpace@Cambridge management groups, obtaining assurances that DSpace@Cambridge will sustain the training and guidance resources created. The DSpace@Cambridge support team plans to undertake moderate continued maintenance and development of the Cambridge data management web pages and is working to further expand and advertise its role as a data management support resource. Further cases for sustainability and expanding support for research data are being made within the context of the repository and will form part of its review processes and development plans. Discipline-specific projects, such as DataTrain (Archaeology, Social Anthropology) will continue to feed into Incremental's Cambridge and public resources.

A case was recently made to the University of Glasgow Digital Preservation Advisory Board (DPAB) to sustain the resources produced by Incremental, and investigate options for additional work to address the other recommendations from the scoping studies. The two main requirements to come out of these studies were: simple guidance that is easy for staff to locate and understand; a preservation infrastructure, encompassing technical provision and hands-on support.

HATII within Glasgow has committed to undertaking basic maintenance of the web pages initially, and the DPAB has agreed to embed these so they become part of central services in the long term.

There has been much success at both institutions in terms of embedding the resources from Incremental. At Cambridge, the Research Office, which has to approve all funding applications, has agreed to direct researchers to the data management web pages and DSpace@Cambridge's support team during the application stage of externally-funded projects. Similarly, at Glasgow, Research and Enterprise has agreed to direct researchers to the pages, and the Digital Preservation Advisory Board is investigating adding a checkbox for data planning to the PAF form (part of the grant process) to flag the requirement and link to relevant guidance and support.

Significantly, many researchers, administrators, and IT professionals at both Cambridge and Glasgow University have agreed to direct students and colleagues to the Incremental web resources, or to link to them

¹⁶ See the Incremental Final Report at: <http://www.lib.cam.ac.uk/preservation/incremental/>

directly. There has been ongoing liaison with University services and support staff to ensure the guidance is accurate and appropriate.

Continued training support will be provided at Cambridge through annual and term library-run training courses, and Incremental will feed into departmental partners' regular academic training courses. At Glasgow, continued training support will be provided through the annual data management course hosted by Staff Development Services and a DCC roadshow is being organised for June 2011 to help the University plan further data infrastructure development.

Project staff have also broken down the person-costs associated with Incremental in a table in the report (Table 1), in the hope that it will assist other institutions planning similar data management support work. Scenarios for sustaining and further enriching Incremental resources at the Universities of Cambridge and Glasgow are presented in another (Table 2)¹⁷.

5.3. POLICY

The MaDAM project at the University of Manchester has developed a pilot infrastructure for the better management of data along the research lifecycle, from data capture to storing, preservation and dissemination. The infrastructure will include a technical (hardware and software) system and curation and governance policies for research data management.

The undertaking is timely not only to support researchers to manage their research data well but also to help them comply with legal and funder policies for better data curation procedures. In talking to a number of data management policy stakeholders at University of Manchester as well as reviewing the general funding requirements, the MaDAM project has produced a Landscape Review document on Policies, Legal & Ethical Perspectives, Stakeholders and Institutional Settings¹⁸. A Data Management Planning component has been included in the infrastructure to provide adequate metadata and guidance for its users. Users have been provided with assistance in the development of Data Management Plans (DMPs), both through a series of workshops, and also with in-tool support (eDMP). The in-tool support consists of a series of help pages which advise the researcher on data management planning, funding body policies, contacts etc, as well as intelligent functions such as automatic review dates. MaDAM has also developed an interface with the Research Office's Research Management Systems to make the completion of DMPs less laborious, allowing us to populate part of the eDMP with information from such systems.

MaDAM is integrated with ongoing strategic initiatives within the University such as the development by the John Rylands University Library of a new strategy for the digital age including eScholar an institutional repository for research outputs. eScholar will be linked with MaDAM and function as the dissemination endpoint of its research data. Furthermore, the Storage, Archiving and Curation project group has produced a proposal for a wider Research Data Management Service (RDMS) at the University of Manchester supported by The University's IS Strategy Board, Manchester Informatics (Mi) and The John Rylands University Library. MaDAM is being used as a demonstrator for this and its findings and infrastructure are instrumental in the envisioned strategy and service provision with the aim to roll out this service incrementally, adding research groups sequentially. This proposal within a wider University research data management strategy is currently being explored. The MaDAM pilot acts as a first step in analysing how a university-wide data management service can be introduced.

¹⁷ These tables are contained in the report on Incremental Business Case available at:

http://www.jisc.ac.uk/whatwedo/programmes/mrd/outputs/benefit_studies

¹⁸ http://www.merc.ac.uk/sites/default/files/MaDAM-BenefitsCaseStudy-FINAL_draft.doc

The JISC funded phase of MaDAM ends in June 2011. Continuation of the service would run from July 2011 onwards and is being addressed by the MaDAM business case to the University.

5.4. TECHNICAL INFRASTRUCTURE AND TOOLS

Business cases have been developed for the FISHNet disciplinary repository (FISHNet project) and the institutional Database as a Service (DaaS) software (as part of the Sudamih project). Sustainability for these is discussed below.

FISHNet

The FISHNet project ran from 1st October 2009 to 31st March 2011, with the objective of developing a repository environment for supporting and promoting:

- Management and curation of datasets in freshwater science;
- Low barriers to depositing datasets;
- Data sharing within the freshwater community.

FISHNet is an example of a subject-based repository developed within the RDMI projects and benefits from the project are discussed in greater detail in section 3.4. The freshwater science community is thinly spread across multiple institutions, and it was apparent that an environment that supported the effective curation and sharing of data would be highly beneficial to those working in the field. The scope of the project therefore covers not only freshwater researchers from Universities, but also other research institutions. As a result of the project, any freshwater scientist registered with the system will be able to share their datasets with the rest of the community in an environment that offers long-term curation facilities, as well as being able to find freshwater datasets shared by other users.

The system developed by the project will be housed at the Freshwater Biological Association and managed in the long-term by the FBA's Data and Information services staff. The provision of data and information has been a key part of the FBA's remit since it was founded in 1929, as one of its charitable objectives was to maintain an information store on the subject of freshwater science. The provision of digital data and information is now a key part of the FBA's long-term business strategy. As the FBA is a membership-based organisation, and its members are distributed among the freshwater community at a variety of institutions, the FBA is uniquely placed to provide a central focus for data sharing among freshwater scientists.

The system delivered at the end of the project is considered to be sufficiently robust for rolling out into production. However, the sustainability of the system after the completion of the project will incur additional costs, for the day-to-day operation of the system, for maintenance of the software, and for enhancements to the software. A business case has been presented to the FBA for this.

The estimated costs in the business case for the long-term support of the repository system are broken down into two broad categories: hardware costs and staff costs.

An initial investment of approximately £30K has been made by the FBA to provide the necessary storage space for the data archive projects in which the FBA is involved. The costs of maintaining hardware and replacing obsolescent hardware are factored into the FBA overheads, as these apply to multiple projects as well as to the general IT services provision for the FBA. As such, hardware is not a direct cost of the FISHNet system. Bug fixing and, to some extent, system enhancements will be implemented as part of the FBA's core business, although for more substantial enhancements additional development funding may be required.

It is envisaged that staff support of the repository will require two distinct roles:

- Repository Manager, who (among other activities) will be involved in the review and QA of deposited data and metadata;
- Software Developer, for fixing bugs in bespoke software and installing security patches and other upgrades to third party software.

The staff FTE required for the Repository Manager role is difficult to estimate at the moment, as it depends on the future uptake of the system among researchers. Initially, FBA are allocating 0.1 FTE, although it could grow to 1.0 FTE (i.e. a full-time post) if uptake grows to the maximum level currently envisaged. The Software Developer is estimated at 0.1 FTE.

Full-Time Equivalent staff costs for each of these roles amount to approximately £50K per annum at Full Economic Cost. Thus initially FBA expect the cost of maintaining the repository to be approximately £10K per annum, although this figure will increase as more users begin using the system.

Sudamih DaaS

The Supporting Data Management Infrastructure for the Humanities Project (Sudamih) is working on two particular aspects of research data management infrastructure: training for researchers; and the development of simple and intuitive 'Databases as a Service' (DaaS) software which researchers can use to create, edit, and share their own research databases. The project is working with researchers from the Humanities Division at the University of Oxford to meet the particular needs of humanities scholars in the first instance. The University intends ultimately to expand the infrastructure developed to the other academic disciplines as well.

Besides the feedback from the research data management workshop, the Sudamih Project has also received useful and valuable advice from the Oxford Roman Economy Project (OXREP) which can be used as a case study¹⁹. The OXREP is an ongoing project at the University of Oxford that is seeking to build a comprehensive (and quite complex) database of sites of economic activity in the Roman World. It is being assembled from a number of existing databases, with data being cleaned, standardized, and added to the new database. The project involves original research and it is intended that the OXREP database will continue to grow and be added to in the future.

Dr. Miko Flohr, the Assistant Director of the OXREP, estimated that if the database work they had conducted during 2010 had used a complete version of the DaaS rather than the Access database which they were in practice working with, they would have saved approximately 21% on staff time, primarily due to simplifications to database structuring and through controlling and standardizing data contributions. He indicated that for other projects the DaaS would be likely to save money by reducing expertise requirements.

The savings made by moving the OXREP to a centrally-hosted Virtual Machine was estimated to be even greater. The IT Officer of the Classics Faculty, which currently hosts the OXREP database and Web front-end on a departmental VM, is planning to move it to a centrally-hosted and maintained VM. The cost savings of so doing, when the staff time needed to look after the VM is taken into account, amount to approximately 37%. Given the economies of scale offered by a centrally-supported VI, this saving may be expected to increase further as the infrastructure is enlarged.

Although obviously every humanities database project will be different, the OXREP hopefully illustrates the kind of savings which could be achieved by projects by switching to a centrally-hosted DaaS service in the future.

¹⁹ <http://oxrep.classics.ox.ac.uk/nw/index.php>

At present, the DaaS is a pilot service. Although the functionality is in place for users to import and export databases, structure, edit, and search databases, the system is not yet robust enough to be used for more than test purposes, nor is it intuitive or user-friendly enough for users to be able to carry out regular activities without step-by-step instructions. The university will therefore need to develop the DaaS further to bring it to the standard expected of a supported, documented, production-ready service for research.

As well as improving the usability of the DaaS, the university shall be conducting a more thorough return on investment analysis and creating a full business plan during the follow-on Virtual Infrastructure with Database as a Service (VIDaaS) Project.²⁰

Quantification of research costs can be difficult. Sudamih's attempts to apply the DAF methodology to data assets created by researchers was complicated by the fact that the researchers themselves were not cost-aware and had no real sense of the value of their work in financial terms. As reported in the Sudamih DAF evaluation²¹:

“in addition to the difficulties in estimating the financial value of researcher time, none of the respondents in this sample made any reference to costs to the institution (for server space, library access, and so forth), despite costs borne by the university or department being specifically mentioned in the question. This suggests that asking researchers alone may not always give a complete picture of the costs of creating data resources.”

When Sudamih came to assessing the costs and benefits of the services it was creating staff found it helpful to ask researchers to estimate the time taken spent performing quite specific and clearly-defined tasks and basing costings on this. It was also important to ask precisely who else had a role in the research tasks undertaken, as researchers tended not to think about the non-academic support they received.

Measuring the benefits of data management can also be difficult. Whilst the potential benefits of implementing research data management tools or training might be easy to enumerate, they can be hard to quantify. Illustrative case studies or the measurement of the minimum impact required to bring a return on investment may have to suffice as adequate alternatives to a thoroughly-costed business plan in such circumstances.

²⁰ <http://vidaas.oucs.ox.ac.uk/>

²¹ Patrick, M., „Use of the Data Audit Framework within the Sudamih Project“, 2010, pp.10-11.
<http://sudamih.oucs.ox.ac.uk/docs/Use%20of%20the%20DAF.pdf>

6. CONCLUSIONS

6.1. DATA MANAGEMENT TRAINING AND GUIDANCE

Better data management can simply help individual researchers improve the efficiency with which they use the information they have gathered. Less time spent searching for notes and sources means improved rates of publication. Better management of data and better connections between sources and notes can also mean better research, as the way researchers manage their information affects the way they recall and use it.

Data management training is just one aspect of the data management infrastructure that needs to be implemented in a university in order to maximize the value of that institution's research data, but it is an important aspect which underpins all other stages of the research data lifecycle. Furthermore, basic researcher training can be implemented before all the other aspects of data management infrastructure are fully established, enabling 'quick wins' in terms of better data management.

In terms of guidance, projects discovered from researchers that existing information on best practice and requirements is uniformly difficult to locate and use on most university websites. Projects have developed streamlined guidance and FAQs on research data management which may be useful exemplars to others.

6.2. RESEARCH DATA POLICY

A range of RDMI projects have focussed on data policy and several highlight the fact that Research Councils UK (RCUK) has recently agreed and issued seven common principles on data policy²². The RCUK common principles on data policy provide an overarching framework for individual Research Council policies on data. RDMI outputs support a number of the RCUK common principles in particular:

- Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property;
- Institutional and project specific data management policies and plans should be in accordance with relevant standards and community best practice. Data with acknowledged long-term value should be preserved and remain accessible and usable for future research;
- To enable research data to be discoverable and effectively re-used by others, sufficient metadata should be recorded and made openly available to enable other researchers to understand the research and re-use potential of the data. Published results should always include information on how to access the supporting data;
- In order to recognise the intellectual contributions of researchers who generate, preserve and share key research datasets, all users of research data should acknowledge the sources of their data and abide by the terms and conditions under which they are accessed; and
- It is appropriate to use public funds to support the management and sharing of publicly-funded research data. To maximise the research benefit which can be gained from limited budgets, the mechanisms for these activities should be both efficient and cost-effective in the use of public funds.

²² <http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx>

6.3. RESEARCH DATA INFRASTRUCTURE

The RDMI projects have successfully demonstrated a range of infrastructure options including disciplinary repositories shared by researchers from across many different institutions; or providing centrally managed IT infrastructure for data management within a university or department. These can provide specialist staff, and leverage specialist research related expertise, providing an effective model for supporting research groups.

However there are many challenges involved in developing shared models. The most significant relate to issues to do with funding and culture change to provide ongoing funding for the development and maintenance of facilities and staff.

6.4. OUTCOMES

In total, the projects have been able to accumulate an impressive body of evidence of the benefits from the programme. There are already significant developments and lessons which will be of interest to a wide range of institutions and funders.

Assessing benefits and impact from research data management is still novel and often very challenging for projects particularly if they are of relatively short duration. However a case study approach supported by templates and guidance on best practice proved valuable and was successfully applied by all 8 RDMI projects.

The projects were on average of 18 months duration so this report and the project work on which it is based, should be seen as a formative assessment of those benefits. The full impact of the projects may only be apparent and measured over a longer timespan.

6.5. THE CHALLENGES IN DEFINING METRICS FOR RESEARCH DATA MANAGEMENT PROJECTS

As mentioned at various points in the report, projects faced many challenges in measuring the scale of identified benefits from their work. These included:

Data Citation

If a greater return on investment for data produced by publicly funded research is to be achieved, and data is to be more effectively shared and reused, researchers and data curators should receive greater recognition for the effort involved in making high quality data available. This, in turn, requires that data should be more readily citable. The necessary conventions and digital identifiers are largely in place (although challenges remain around complex, dynamic and aggregated datasets). The real barriers are conventions and practice around citation which prioritise secondary works (articles etc), tend not to include data in the list of references, or limit the length of such lists. Journals, scholarly societies and researchers should be encouraged to include citations of data in the references section of journals articles. These current limitations clearly mitigate the application of data citation metrics in this important area. Furthermore any changes may take considerable time to reach a critical mass of adoption and citations.

Timescales and Attribution

The projects are of short duration typically only 18 months and pilot services are brought on stream when well into the projects. Hence measurements of benefits are challenging. Many benefits will only emerge over a longer timescale. Longer-term effects tend to arise from a complex combination of developments and circumstances which can be difficult or impossible to disentangle and attribute to use of a single data repository, dataset, or research data management project. Projects therefore often attempted to gather what information they can about more immediate short-term benefits, on the principle that if no short-term benefit can be detected the likelihood of significant long-term benefits is also doubtful.

Maturity and Critical Mass

Usage of repositories and acceptance of data sharing often grows very slowly at first and may only become significant when a critical mass of data has been assembled. The length of time before or the point in time when a repository or dataset's impact is assessed can therefore be critical.

If a discipline or sub-discipline or its data management is still relatively immature, the focus of repositories activities may be on influencing its community and developing common standards and practice. Data use may still be relatively difficult and low until these are established.

Resources and Skills

The application of the KRDS Cost Model and activity-based costing proved too challenging for most projects: only the I2S2 project implemented this approach, largely because the resources required to use activity-based costing were not available in most projects and it is difficult to apply without previous experience. Where cost information exists (e.g. staff timesheets) it may still need significant time for processing and verification before it is in a form which can be used for comparison and analysis. However projects can learn from and apply the "rules of thumb" and key lessons concerning activity costing from comparable activities and projects. Development of the KRDS User Guide and KRDS Factsheet during the programme should assist in this process.

Feedback on the use of DAF and DRAMBORA by the RDMI projects has revealed that these assessments can be quite arduous and often require considerable effort to carry out effectively. Feedback also indicated that users are looking for assessments that result in practical recommendations for improvement rather than simply identifying where they currently stand. These insights are informing the development of the new CARDIO dataset.

Scope

A lesson learned from the applications of AIDA, DAF and DRAMBORA is that starting with a small, well-defined scope is more successful than starting on a broader scale. So, looking at particular projects, research group activities, or department level activity is generally more fruitful than looking at the institution as a whole. Feedback also informed us that the end result of the assessment is often less valuable than the actual assessment process itself. In most cases the wide range of stakeholders involved in these assessments (researchers, central support, IS) had little contact with each other on a day to day basis and didn't always understand each other's concerns and requirements.

6.6. ASSESSING BENEFITS AND METRICS IN FUTURE PROGRAMMES

A number of successes can be identified in the programme for future use elsewhere. Identification of benefits using the KRDS Benefits Framework was successfully applied by a range of RDMI projects and their feedback has provided enhancements to the KRDS methodology and documentation. I2S2 has also added a Value Chain and Impact Tool. These will benefit future programmes and projects. Similarly, the feedback received on the applications of the individual assessment tools such as AIDA tools has helped to shape the development of the Collaborative Assessment of Research Data Infrastructure and Objectives (CARDIO) tool, which will be available for future projects.

Finally the use of templates and best practice guidance to support the RDMI project in preparing their benefits case studies proved to be a successful approach. However further work is needed on definition and support for relevant metrics to further enhance its utility to future programmes.

7. FURTHER INFORMATION

7.1. MRD PROGRAMME

JISC Managing Research Data Programme Web Pages:

<http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx>

7.2. RDMI PROJECTS

ADMIRAL Project Website: <http://imageweb.zoo.ox.ac.uk/wiki/index.php/ADMIRAL>

Blueprint (IDMB) Project Website: <http://www.southamptondata.org/>

FISHNet Project Website: <http://www.fishnetonline.org/>

HALOGEN Project Website: <http://www2.le.ac.uk/offices/itservices/resources/cs/ps0/project-websites/halogen>

I2S2 Project Website: <http://www.ukoln.ac.uk/projects/I2S2/>

Incremental Project Website: <http://www.lib.cam.ac.uk/preservation/incremental/>

MaDAM Project Website: <http://www.merc.ac.uk/?q=MaDAM>

Sudamih Project Website: <http://sudamih.oucs.ox.ac.uk/>

7.3. MRD PROGRAMME SUPPORT PROJECTS

The Cost/Benefit and Business Case Support Project webpage: <http://www.beagrie.com/dmi.php>

The Integrated Data Management Planning Toolkit and Support Project website:

<http://www.jisc.ac.uk/whatwedo/programmes/mrd/supportprojects/idmpsupport.aspx>

7.4. KEEPING RESEARCH DATA SAFE REPORTS AND TOOLS

Keeping Research Data Safe (KRDS1) - Final Project Report (2008):

<http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx>

Keeping Research Data Safe 2 (KRDS2) - Final Project Report (2010):

<http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx#downloads>

Keeping Research Data Safe 2 (KRDS2) Project website - provides access to the costs data survey results and other supplementary materials: <http://www.beagrie.com/jisc.php>

KRDS Factsheet - (PDF File) - This A4 four-page factsheet is a concise summary of KRDS's key findings:

http://www.beagrie.com/KRDS_Factsheet_0910.pdf

KRDS User Guide (PDF File) - The KRDS User Guide is an edited selection and synthesis of the guidance in the KRDS reports combined with newly commissioned text and illustrations:

http://www.beagrie.com/KeepingResearchDataSafe_UserGuide_v1_Dec2010.pdf

KRDS Tools (PDF Files) - The KRDS activity cost model is available to download in two versions together with a KRDS Benefits Taxonomy (note guidance on the use of the activity models and benefits taxonomy is available in the KRDS User Guide):

- KRDS2 Activity Model "Lite" - a one page overview of the main phases and activities in the model: http://www.beagrie.com/KRDS2_Activity_Model_lite.doc

- KRDS2 Activity Model "Detailed" - the full model (12 pages) with all definitions and sub-activities listed: http://www.beagrie.com/KRDS2_Activity_Model_detailed.doc
- The KRDS Benefits Taxonomy - Version 2.0 (Dec 2010) a summary table of the KRDS benefits taxonomy developed in KRDS2 and updated for the KRDS User Guide: <http://www.beagrie.com/KRDSBenefitsTaxonomyv2.doc>

The KRDS Digital Preservation Benefit Analysis Tools Project is developing a toolset from KRDS tools/an I2S2 Value-Chain and Impact Tool (for released on project website summer 2011): <http://beagrie.com/krds-i2s2.php>

7.5. CARDIO AND OTHER TOOLS

CARDIO (Collaborative Assessment of Research Data Infrastructure and Objectives) Tool:

<http://cardio.dcc.ac.uk/>

CARDIO integrates the concepts of the aforementioned tools rather than their workflows which can be complex and time consuming to undertake. CARDIO draws together the broad concepts of risk, maturity and capacity at both data and infrastructural levels and serves as a jumping off point for those wishing to undertake more detailed DAF or DRAMBORA assessments.

AIDA (Assessing Institutional Digital Assets): <http://aida.da.ulcc.ac.uk/wiki/index.php/AIDA>

DAF (Data Asset Framework): <http://www.dcc.ac.uk/resources/tools-and-applications/data-asset-framework>

DRAMBORA (Digital Repository Audit Method Based on Risk Assessment):

<http://www.dcc.ac.uk/resources/tools-and-applications/drambora>