JISC

Project Document Cover Sheet

Project Information			
Project Acronym	I2S2		
Project Title	Infrastructure for Integration in Structural Sciences		
Start Date	1st Oct 2009	**End Date**	30th June 2011
Lead Institution	Universities of Bath and Southampton		
Project Director	Liz Lyon (UKOLN)		
Project Manager & contact details	Manjula Patel 01225 386547; m.patel@ukoln.ac.uk		
Partner Institutions	Universities of Bath, Southampton, Cambridge; STFC; Charles Beagrie Ltd.		
Project Web URL	http://www.ukoln.ac.uk/projects/I2S2/		
Programme Name (and number)	Managing Research Data (Research Data Management Infrastructure)		
Programme Manager	Simon Hodson		

Document Name			
Document Title	I2S2 Business Models & Sustainability Plan		
Reporting Period	*N/A*		
Author(s) & project role	Neil Beagrie + Project Partners		
Date	22nd June 2011	**Filename**	I2S2-BusinessModels&SustainabilityPlan-110622
URL			
Access	X Project and JISC internal	☐ General dissemination	

Document History		
Version	**Date**	**Comments**
1.0	23/05/2011	Draft circulated for partner comments and contributions Ownership of document transferred to Manjula Patel
2.0	06/06/2011 07/06/2011 09/06/2011 15-17/06/2011	Added in contributions from Liz Lyon and Simon Coles Added in contributions from Neil Beagrie Added in contributions from Brian Matthews Added sections 2.7, 4.6 and completed Exec Summary Circulated to Partners for final comments
3.0	22/06/2011	Final version submitted to JISC

I2S2 BUSINESS MODELS AND SUSTAINABILITY PLAN

VERSION 3.0, 22ND JUNE 2011

CONTENTS

EXECUTIVE SUMMARY

There is currently a functional chasm between researchers working in their home institution and at centralised facilities such as Diamond and ISIS. Researchers need to move data across institutional and domain boundaries in a seamless and integrated manner. The Infrastructure for Integration in Structural Sciences (I2S2) project has attempted to "bridge the chasm" and develop a robust data infrastructure to enable these seamless transformations to take place routinely and to greatly increase researcher efficiency and productivity. There is also likely to be greater return on investment in the central facilities such as Diamond through more cost-effective use of resources by the client base.

This Document sets out proposals and business models for sustaining the work of I2S2 beyond the life of the initial project which ends in June 2011.

The project has aspired to bring about significant benefits which are quantifiable, sustainable and transferable to the entire structural science domain as well as across other disciplines and across institutions. More specifically, the harmonisation of distributed representations of data models through abstraction into an Integrated Information Model which underpins research across multiple sites and global locations is a significant step forward within the structural science community. It will facilitate data validation, data sharing, data access and management and data preservation in the longer term. In addition, it is hoped that key principles and lessons learnt will be transferable into domains such as materials science and engineering.

We examine the strategies and policies of RCUK, EPRSC, JISC and the Digital Curation Centre (DCC), STFC and its facilities, and the UK Research Integrity Office's Code of Practice for Research and discuss how the work of the I2S2 project and its future implementations are closely aligned with them.

We also discuss a range of substantial benefits in terms of research effectiveness and research efficiency that we have been able to identify from I2S2 and potential metrics to measure their future impact. Two detailed case studies are provided describing benefits from a researcher's and a service provider's perspective.

The sustainability issues in I2S2 are complex since the project has a disciplinary community focus and also spans multiple organisations rather than being an initiative within a single institution. The business case for the continuation of the I2S2 work is predicated on an Integrated Service approach which delivers a suite of joined-up services derived from the harmonisation of existing services at local (e.g. institutional laboratory), national (e.g. National Crystallography Service) and international (e.g. STFC) levels. An approach based on integration has the advantage of improving the probability that interventions developed in I2S2 become fully embraced and embedded into the pre-existing infrastructure. However, to fully realise the vision of the I2S2 Project, it would be necessary to undertake a second phase of the project with the aim of completing implementation of the pilot infrastructures; development of ICAT-Personal into a robust research data management tool; further development of the I2S2 Information Model; completion of the cost-benefits analysis and impact work; and extended training of the target communities identified in the project. A second phase of the project would allow coordinated action on all of these fronts.

Given the uncertainty of appropriating funding for a follow-on implementation phase of I2S2, we have examined a series of options for sustaining key outcomes from the project; we therefore offer the following conclusions and recommendations:

- I2S2 Information Model will be further developed in current projects such as PaN-Data and the UMF Smart Research Framework, which will sustain and promote the model over the next 2-3 years, and build tools which will use and develop it further. If the Model is widely adopted, we will seek to support it through an open source community effort.
- The ICAT-Personal Tool will likely become integrated into the suite of ICAT tools which support the data management needs of ISIS and DLS to enable the data analysis phase of the lifecycle activity model as explored in I2S2.
- The NCS intends to update its data and information management systems with a unified framework, underpinned by the I2S2 Information Model that supports all aspects of facility operation, covering the whole I2S2 research activity lifecycle model.
- Our work on assessing benefits and impact has already led to innovations which are being further developed and sustained via a JISC 15/10 programme project (Digital Preservation Benefit Analysis Tools) involving a range of partners and data services. The tools will be user tested, documented, made freely available, promoted by a range of services, and have value-added support via consultancy if required. This will allow independent support for and evolution of this work.
- The I2S2 community is only one small part of the much wider structural science community, and a co-ordinated programme of awareness-raising, training, professional development and networking across all the structural science domains, is required; these might include physics, mineralogy, earth sciences and some aspects of bio-engineering. Extending the I2S2 approach beyond the chemistry domain is a priority.
- In the medium-term we will continue to disseminate the results of I2S2 as well as KRDS/I2S2 Benefit Analysis tools and advocate their use at workshops, conferences and disciplinary meetings.
- Knowledge transfer to equipment and instrument manufacturers; the NCS has been in close collaboration with the instrument provider, Rigaku, regarding information and data management and a plan has been drawn up to incorporate elements of the I2S2 Information Model into the Rigaku data management framework. This will involve addition of I2S2 elements into the Rigaku SIMS (Sample Information Management System) – an XML description of samples and their attributes which will enable data management from a facility perspective (which was not previously possible) and will be rolled out with all Rigaku software in the future.

In section 5 we provide initial estimates of cost and timescales for continuation of the primary strands of work as outlined above.

1. PROJECT DESCRIPTION

The Infrastructure for Integration in Structural Sciences (I2S2) project has been identifying requirements for a data-driven research infrastructure in 'Structural Science', focusing primarily on the domains of Chemistry and Crystallography.

The project has addressed three complementary infrastructure "axes" as shown in Figure 1:

- Scale and complexity: from small laboratory equipment through institutional installations to large scale facilities such as the DLS and ISIS at STFC;
- Inter-disciplinary: research across domain boundaries;
- Data lifecycle: time-factored data flows and data transformations.

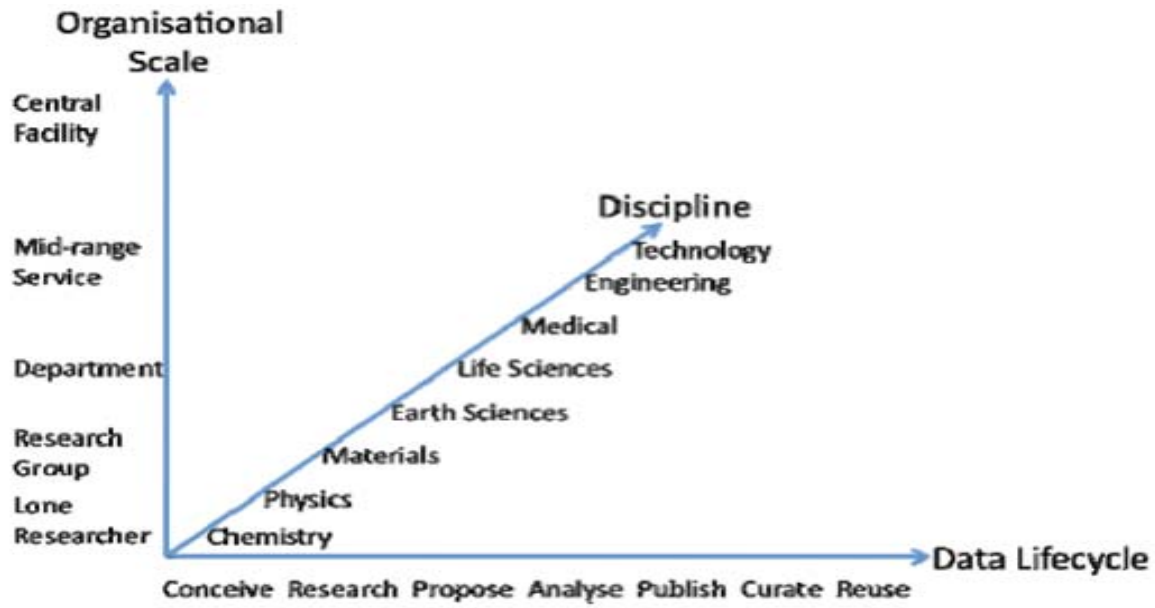

Figure 1*: Three Dimensions of Research Data Management Infrastructure*

It should be noted that a fourth dimension, human curation infrastructure i.e. skills development and training aspects, had been recognised by the team but was not in the scope of this project. Similarly large-scale software and policy development, whilst relevant to a data management infrastructure, were not under consideration in this project.

The DLS is a third generation synchrotron radiation source operated by Diamond Light Source Ltd, a company co-owned by STFC. ISIS is a world-leading pulsed neutron and muon source operated by the STFC at Rutherford Appleton Laboratory. Both of these facilities enable scientists to investigate the structure and dynamics of matter, such as biological tissues, polymers and catalysts, at the atomic and molecular level. The facilities are used by thousands of scientists internationally in a wide range of scientific disciples, covering topics at the forefront of Physics, Chemistry, Materials Science, Earth Science, Engineering and Biology. Simon Coles at the NCS (Soton, Crystallography) makes regular use of DLS, whilst Martin Dove (Cambridge, Earth Sciences) is a major user of ISIS.

During the first phase of the I2S2 project, we commissioned a comprehensive data management requirement report[1] for the structural science research arena. The major findings from the report were:

- The four broadly defined levels of research science examined in the report (individual researcher, research team, medium-level service, and large-scale facility) revealed the huge diversity of requirements depending on the situation, circumstances and level of data management infrastructure currently in place;
- At present individual researchers, groups, departments, institutions and service facilities appear to be all working within their own technological frameworks so that proprietary and insular technical solutions have been adopted (e.g. use of multiple and/or inconsistent identifiers). This makes it onerous for researchers to manage their data which can be generated, collected and analysed over a period of time, at multiple locations and across different collaborative groups. Researchers need to be able to move data across institutional and domain boundaries in a seamless and integrated manner.

The implementation plan[2] for the I2S2 project was written after we had gained some initial experience of designing and developing a preliminary pilot implementation for capturing, storing, and visualising the derived data generated throughout the analysis pipeline of an exemplar structural science experiment. It narrowed down our efforts to a few key areas that need most attention. Specifically, we addressed six out of sixteen findings resulted from requirements gathering process namely:

- A robust data management infrastructure which supports each researcher in capturing, storing, managing and working with all the data generated during an experiment;
- Internal sharing of research data amongst collaborating scientists, such as between a PhD student and supervisor;
- Capture, management and maintenance of:

  (1) Metadata and contextual information (including provenance);

  (2) Control files and parameters;

  (3) Versioning information;

  (4) Processing software;

  (5) Workflow for a particular analysis;

  (6) Derived and results data;

  (7) Links between all the datasets relating to a specific experiment or analysis.
- Changes should be easily incorporated into the scientist's current workflow and be as un-intrusive as possible;
- It was clear that the processing pipeline in many scientific experiments tend to be near digital, relying on suites of tools, applications software and very often

[1] http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2-WP1-D1.1-RR-Final-100707.pdf

[2] http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2-WP3-D3.2-ImplementationPlan.pdf

customised software. There is therefore a need to document, maintain and curate such software and support its future development;

- The Core Scientific Metadata Model (CSMD) and its implementation in the ICAT database of the Science and Technology Facilities Council (STFC) is a good candidate for further development and extension to take account of the needs of organisations outside of the STFC.

2. STRATEGIC ALIGNMENT

The work of I2S2 is closely aligned to the strategies of a range of major stakeholders in the project including RCUK, EPSRC, JISC, DCC, STFC and the Diamond and ISIS facilities as detailed below. In summary, the I2S2 project enhances the scientific process through improvements to the managing and sharing of research data. Good research data management practice allows reliable verification of results and permits new and innovative research to be built on existing information. The project has also worked to identify and measure associated benefits. This is important if the full value of public investment in research is to be realised.

2.1 RESEARCH COUNCILS' COMMON PRINCIPLES ON DATA POLICY

Research Councils UK (RCUK) has recently agreed and issued seven common principles on data policy[3]. Making research data available to users is a core part of the Research Councils' remit and is undertaken in a variety of ways. The RCUK common principles on data policy provide an overarching framework for individual Research Council policies on data policy. I2S2 outputs support a number of the RCUK common principles in particular:

- 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; 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.

2.2 JISC AND THE DIGITAL CURATION CENTRE (DCC)

The JISC has made significant investments in research data management in the UK higher education sector. From 2004, it has funded the Digital Curation Centre, which provides advocacy, tools and resources to support research data management within higher education institutions. The University of Bath is a DCC partner and the I2S2 Project is one of a series of data-focussed research and development projects supported by the DCC. The

[3] http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx

DCC has also developed a number of tools such as the Data Asset Framework and DMPOnline, which aim to assist institutions in the assessment of and planning for effective data curation.

The JISC has also developed a major innovation programme around research data management, which includes a series of funded innovation projects exploring aspects of data infrastructure (within institutions and selected domain communities such as structural science), data publishing, data citation, training and cost-benefits. In the current challenging economic climate, the policy and funding drivers for institutions and research communities alike, to adopt cost-efficient processes, to articulate benefits and demonstrate value associated with public investments in research, are particularly pertinent in this context; I2S2 is positioned at the centre of this space.

2.3 EPSRC Policy Framework on Research Data Sets

The EPSRC Policy Framework on Research Data sets[4] out EPSRC's expectations concerning the management and provision of access to EPSRC-funded research data. The framework was endorsed by the EPSRC Council in March 2011 and implemented from 1st May 2011. The expectations arise from seven core principles, which align with the core RCUK principles on data sharing. The policy reflects the principal UK legal provisions intended to assure public access to publicly held information, the most relevant of which to EPSRC-funded research data are contained in the Freedom of Information Act (2000) and the Freedom of Information (Scotland) Act (2002) (Other relevant legislation includes the Data Protection Act 1998, the Environmental Information Regulations 2004 and the Environmental Information (Scotland) Regulations 2004).

EPSRC has the following clear expectations of organisations in receipt of EPSRC research funding - those particularly supported by I2S2 are highlighted below:

i. Research organisations will promote internal awareness of these principles and expectations and ensure that their researchers and research students have a general awareness of the regulatory environment and of the available exemptions which may be used, should the need arise, to justify the withholding of research data.

ii. Published research papers should include a short statement describing how and on what terms any supporting research data may be accessed.

iii. Each research organisation will have specific policies and associated processes to maintain effective internal awareness of their publicly-funded research data holdings and of requests by third parties to access such data; all of their researchers or research students funded by EPSRC will be required to comply with research organisation policies in this area or, in exceptional circumstances, to provide justification of why this is not possible.

iv. Publicly-funded research data that is not generated in digital format will be stored in a manner to facilitate it being shared in the event of a valid request for access to the data being received (this expectation could be satisfied by implementing a policy to convert and store such data in digital format in a timely manner);

[4] http://www.epsrc.ac.uk/about/standards/researchdata

v. **Research organisations will ensure that appropriately structured metadata describing the research data they hold is published (normally within 12 months of the data being generated) and made freely accessible on the internet; in each case the metadata must be sufficient to allow others to understand what research data exists, why, when and how it was generated, and how to access it. Where the research data referred to in the metadata is a digital object it is expected that the metadata will include use of a robust digital object identifier (For example as available through the DataCite organisation - http://datacite.org).**

vi. Where access to the data is restricted the published metadata should also give the reason and summarise the conditions, which must be satisfied for access to be granted. For example 'commercially confidential' data, in which a business organisation has a legitimate interest, might be made available to others subject to a suitable legally enforceable non-disclosure agreement.

vii. **Research organisations will ensure that EPSRC-funded research data is securely preserved for a minimum of 10-years from the date that any researcher 'privileged access' period expires or, if others have accessed the data, from last date on which access to the data was requested by a third party; all reasonable steps will be taken to ensure that publicly-funded data is not held in any jurisdiction where the available legal safeguards provide lower levels of protection than are available in the UK.**

viii. **Research organisations will ensure that effective data curation is provided throughout the full data lifecycle, with 'data curation' and 'data lifecycle' being as defined by the Digital Curation Centre. The full range of responsibilities associated with data curation over the data lifecycle will be clearly allocated within the research organisation, and where research data is subject to restricted access the research organisation will implement and manage appropriate security controls; research organisations will particularly ensure that the quality assurance of their data curation processes is a specifically assigned responsibility;**

ix. **Research organisations will ensure adequate resources are provided to support the curation of publicly-funded research data; these resources will be allocated from within their existing public funding streams, whether received from Research Councils as direct or indirect support for specific projects or from Higher Education Funding Councils as block grants.**

In summary the applicability of these principles to I2S2 are associated with the long-term curation of and access to, research data and specifically note the importance of making appropriate metadata available for discovery purposes.

2.4 Data Policies of the Science and Technology Facilities Council

The major STFC facility the ISIS spallation neutron source (ISIS) has published data policies[5]. To summarise, the main points of the policy of most relevance to I2S2 are as follows:

[5] http://www.isis.stfc.ac.uk/user-office/data-policy11204.html

- All raw data and the associated metadata obtained as a result of free (non-commercial) access to ISIS, reside in the public domain, with ISIS acting as the custodian.
- Access to raw data and metadata beyond the period that it is stored on instrument-related computers will be via a searchable on-line catalogue.
- Access to the on-line catalogue will be restricted to those who register with STFC/ISIS as users of the on-line catalogue.
- Access to raw data and the associated metadata obtained from an experiment is restricted to the experimental team for a period of three years after the end of the experiment. Thereafter, it will become publicly accessible. Any PI that wishes their data to remain 'restricted access' for a longer period will be required to make a special case to the Director of ISIS.
- The on-line catalogue will enable the linking of experimental data to experimental proposals. Access to proposals will only ever be provided to the experimental team and appropriate STFC staff, unless otherwise authorized by the PI.
- Ownership of all results derived from the analysis of the raw data is determined by the contractual obligations of the person(s) performing the analysis.
- ISIS undertakes to provide facilities for the capture of such metadata items that are not automatically captured by an instrument, in order to facilitate recording the fullest possible description of the raw data.
- Researchers who aim to carry out analyses of raw data and metadata which are publicly accessible should, where possible, contact the original PI to inform them and suggest a collaboration if appropriate
- PIs and researchers who carry out analyses of raw data and metadata are encouraged to link the results of these analyses with the raw data / metadata using the facilities provided by the on-line catalogue. Furthermore, they are encouraged to make such results publicly accessible.
- References for publications related to experiments carried out at ISIS must be deposited in the STFC e-Pubs system http://epubs.cclrc.ac.uk/ within six months of the publication date, or during any new application for beamtime, whichever is the earlier.

Other international facilities including the Diamond Light Source (DLS) are drafting or considering similar data policies; a common framework for such data policies is being promoted in the PaNData Strategic Working Group project[6].

2.5 The UKRIO Code of Practice for Research

The UK Research Integrity Office's (UKRIO) Code of Practice for Research[7] has been designed to encourage good conduct in research and help prevent misconduct, in order to

[6] http://www.pan-data.eu/imagesGHD/08/PaN-data-D2-1.pdf

[7] http://www.ukrio.org/sites/ukrio2/the_programme_of_work/code_of_practice_for_research.cfm

assist organisations and researchers to conduct research of the highest quality. It provides general principles and standards for good practice in research, applicable to both individual researchers and to organisations that carry out, fund, host or are otherwise involved in research. Since the publication of the Code in 2009, it has been adopted and used by many research organisations and endorsed by research funders and other bodies.

UKRIO is hosted by Universities UK and has the support of a number of UK organisations with interests in research including: the four UK Departments of Health, the four UK Higher Education Funding Councils, the Academy of Medical Sciences, the Association of the British Pharmaceutical Industry, the Association of UK University Hospitals, the Biotechnology and Biological Sciences Research Council, the Committee on Publication Ethics, the General Medical Council, the Medical Research Council, the Medical Schools Council, the Medicines and Healthcare products Regulatory Agency, Research Councils UK, the Royal College of Physicians, the Royal College of Physicians of Edinburgh, the Royal Society, Universities UK and research charities including the Wellcome Trust.

Implementation of I2S2 will assist organisations in meeting the following aspects of the Code relating to collection and retention of data:

> 3.12.1 Organisations and researchers should comply with all legal, ethical, funding body and organisational requirements for the collection, use and storage of data, especially personal data, where particular attention should be paid to the requirements of data protection legislation. They should also maintain confidentiality where undertakings have been made to third parties or to protect intellectual property rights. Organisations and researchers should ensure that research data relating to publications is available for discussion with other researchers, subject to any existing agreements on confidentiality.
>
> 3.12.2 Data should be kept intact for any legally specified period and otherwise for three years at least, subject to any legal, ethical or other requirements, from the end of the project. It should be kept in a form that would enable retrieval by a third party, subject to limitations imposed by legislation and general principles of confidentiality.
>
> 3.12.5 Organisations should have in place procedures, resources (including physical space) and administrative support to assist researchers in the accurate and efficient collection of data and its storage in a secure and accessible form.
>
> 3.12.6 Researchers should consider how data will be gathered, analysed and managed, and how and in what form relevant data will eventually be made available to others, at an early stage of the design of the project.
>
> 3.12.7 Researchers should collect data accurately, efficiently and according to the agreed design of the research project, and ensure that it is stored in a secure and accessible form.

2.6 Research Benefit and RCUK Impact Strategy

Impact is defined by Research Councils UK (RCUK) as the demonstrable contribution that excellent research makes to society and the economy; it embraces all the extremely diverse ways in which research-related knowledge and skills benefit individual, organisations and nations by:

- Fostering global economic performance, and specifically the economic competitiveness of the United Kingdom;
- Increasing the effectiveness of public services and policy;
- Enhancing quality of life, health and creative output.

RCUK's impact strategy[8] was launched in March 2010 and will ensure that the Research Councils will build on their past successes and maximise and celebrate the impact generated from the research, people and facilities that they fund. The impact strategy has the following aims:

- Engaging key stakeholders;
- Maximising research impact;
- Delivering highly skilled people.

I2S2 supports the aims of the RCUK Impact strategy by engaging with key stakeholders across the physical sciences domain and enhancing research data and processes that contribute to training in critical research data management skills within the scope of the project. Our work on assessing benefits and impact will also be valuable to the RCUK assessment of impact. However, it must be noted our aims in terms of assessing impact are wider than the narrower definition of RCUK as we have included the effect of better research data management on the process of research itself (as noted above this supports RCUK's other common data principles that these activities should be both efficient and cost-effective in the use of public funds).

2.7 Conclusions

The sustainability issues in I2S2 are complex since the project has a disciplinary community focus and also spans multiple organisations rather than being an initiative within a single institution. Consequently, the business case for the continuation of the I2S2 work is predicated on an **Integrated Service** approach which delivers a suite of joined-up services derived from the harmonisation of existing services at local (e.g. institutional laboratory), national (e.g. National Crystallography Service) and international (e.g. STFC) levels. This approach is presented in Figure 2. An approach based on integration has the advantage of improving the probability that interventions developed in I2S2 become fully embraced and embedded into the pre-existing infrastructure.

[8] http://www.rcuk.ac.uk/kei

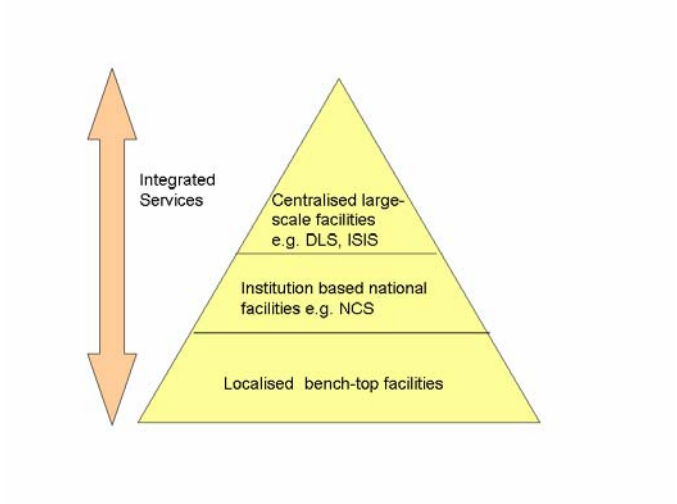

Figure 2*: I2S2 Integrated Service Approach to Sustainability*

The I2S2 Idealised scientific research activity lifecycle model[9] (Figure 3) provides an over-arching framework for the integration and harmonisation of existing services as well as all aspects of the work undertaken in the project. It addresses and supports the recently published and emerging policies of RCUK, JISC/DCC, EPSRC, STFC and UKRIO in the effective and efficient management of research data, its publication, discovery, access and reuse by third parties. This model identifies six broad Data Service Functions: *initiation; collection; analysis; publication; curation* and *discovery*. The policies and strategies examined in the above sections relate to and cover each of these areas:

Initiation: This function includes the development of data management policies and plans in accordance with relevant standards and community best practice and is addressed by JISC/DCC, EPSRC and UKRIO.

Collection: Under this part of the lifecycle model, a physical science experiment is conducted and raw or primary data is collected; this is subsequently processed into derived data upon which various analyses are performed. In I2S2 we found that primary and derived data is rarely accessible by third parties. RCUK, JISC/DCC, EPSRC, STFC and UKRIO all provide guidance with regard to such data.

Analysis: Data resulting from several iterations of analysis tends to be more commonly shared and made available to other research scientists than primary data. RCUK, JISC/DCC, EPSRC, STFC and UKRIO all promote effective management of analysed or results data.

Publication: Effective publication and citation of research datasets requires comprehensive metadata for their interpretation by third parties as we well as persistent identifiers and the JISC/DCC, EPSRC and STFC all endorse the availability of such contextual information.

Curation: The central stack in the lifecycle model addresses the long-term accessibility of scientific research data incorporating the provision of contextual information, *Preservation*

[9] http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2-ResearchActivityLifecycleModel-110407.pdf

Description Information and *Representation Information* as defined in the Open Archival Information System (OAIS) Reference Model[10]. We found that the systematic management of research data for long-tem access is an area that is underdeveloped and requires particular attention. RCUK, EPSRC, STFC and UKRIO all address this part of the lifecycle model.

Discovery: Reuse of research data is very much dependent on it being discoverable and accessible, which in turn is dependent on its being published with rules regarding IPR, embargo and access control being in place. RCUK, JISC/DCC, EPSRC, STFC and UKRIO are all concerned with this part of the lifecycle model.

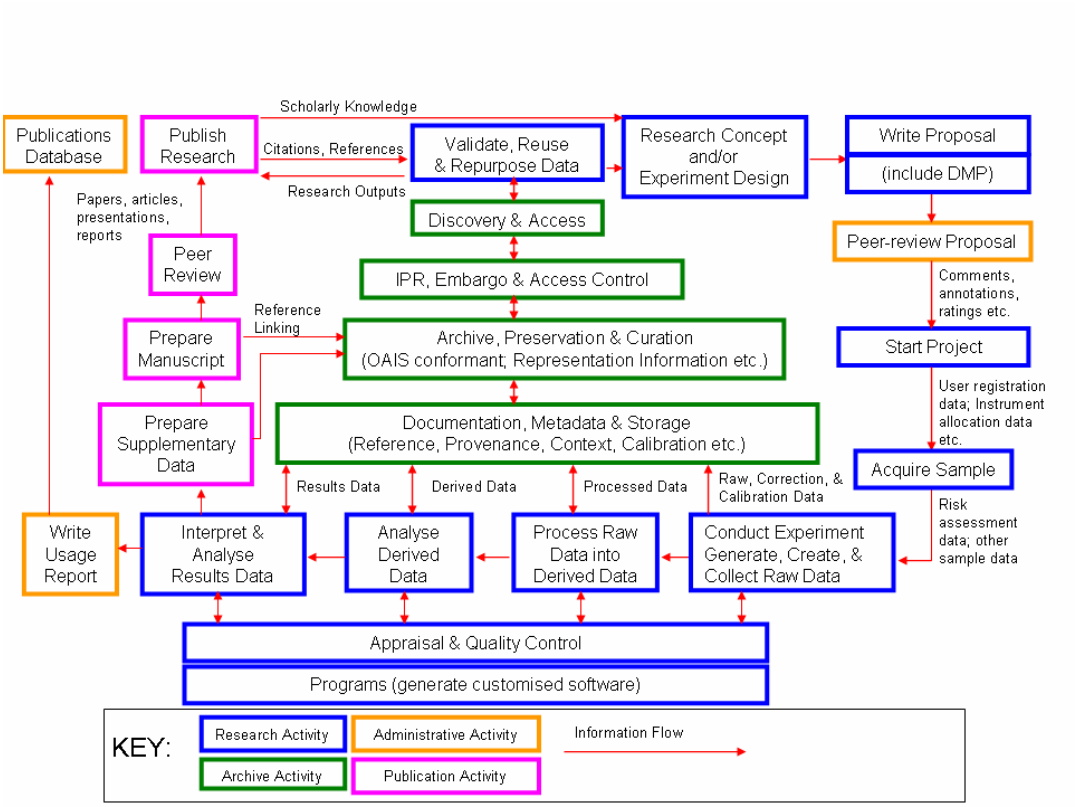

Figure 3: *An idealised scientific research activity lifecycle model*

[10] Consultative Committee for Space Data Systems, Reference Model for an Open Archival Information System, ISO:14721:2002, 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf#search=%22OAIS%20model%22

3. BENEFITS APPRAISAL

A key aim of I2S2 has been to identify the costs and benefits of the integrated approach to information management across local institutional and national facilities proposed by the project. Two parallel benefits cases have explored the perspectives of "scale and complexity" and "research discipline" throughout the data lifecycle and the complementary but often different perspectives of researchers and central facilities on potential benefits.

The I2S2 Benefit Case Study 1 (Service Perspective) was prepared by Simon Coles (National Crystallography Service, University of Southampton) and Neil Beagrie (Charles Beagrie Ltd). It is based on the National Crystallography Service and its interaction with institutional and other central national facilities and how this may be improved by I2S2. It traverses administrative boundaries between institutions and address issues of scale (local laboratory to mid-range national facility to national Diamond synchrotron) and provides a central service perspective of benefits.

I2S2 Benefit Case 2 (Researcher Perspective) was prepared by Martin Dove (University of Cambridge) and Neil Beagrie (Charles Beagrie Ltd). It is based on the research projects of Prof Martin Dove, University of Cambridge using the STFC ISIS central facility. It applies the approach to Mineral Sciences and interactions between individuals, collaborative research groups and facilities, and provides a researcher's perspective of benefits from changes proposed within I2S2.

Each benefits case study has been able to draw on more detailed source material in two benefits use cases[11] prepared as cost/benefit deliverables for the project.

3.1 BENEFITS IDENTIFIED

The primary or major benefits of implementing I2S2 identified by the 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 by these workers;
- **Rapid access to results 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. Implementation of I2S2 can reduce a one-day latency of data access down to five minutes for these researchers;
- **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

[11] http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2_BenefitUseCases_final.pdf

as less significant or have lower impact at the level of individual researchers where this level of scaling does not apply;

- **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.** There is a need to enhance community capability and build additional capacity. 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. 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;
- **Wider access and use**. Facilities will benefit by providing access to results and derived data as part of their services to researchers. 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. This in turn will lead to a benefit for the scientific disciplines;
- **Reducing risk**. There will be less likelihood of introducing error into 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 Pair Distribution Function (PDF) data. Researchers in Benefits Case Study 2 are collaborating with one, and for this company the benefits from I2S2 will be similar to those outlined above 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. Similarly, the NCS has been in close collaboration with the instrument provider, Rigaku, regarding information and data management and a plan has been drawn up to incorporate elements of the I2S2 Information Model into the Rigaku data management framework. This will involve addition of I2S2 elements into the Rigaku SIMS (Sample Information Management System) – an XML description of samples and their attributes. This will enable data management from a facility perspective (not

previously possible with this software) and will be rolled out with all Rigaku software in the future.

3.2 Potential Metrics Identified

Service productivity and efficiencies. I2S2 has developed an activity model of the scientific research data lifecycle and associated tasks[12]. 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. These have been documented in I2S2 Benefits Use Case 1[13]. It should therefore be possible to calculate any work efficiencies and time savings after full implementation of I2S2. As noted above time savings and higher throughput are particularly significant benefits for services because of economies of scale effects they can have dealing with many individual researchers, samples, and experiments. At the same time, 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 can be established.

Extending, training, and self-starting the user community. In Benefits Case Study 2, the wider user community is currently relatively small, producing about 15 papers a year. 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 (Journal of Physics: Condensed Matter 19, 335218, 2007). As of the time of writing, citations number 38 (15 a year for the past two years; we include self-citations here because many of the self-citations are in collaboration with new users, which is a typical trend for any method when new users need help). A clear metric of success here will be an increase in the citation rate of this paper.

New users based on ISIS are inevitably going to grow slowly, because beam time on ISIS instruments is limited. On the other hand, there are potential gains in the user base to be found in the use of new neutron facilities (e.g. at reactor sources on specialised instruments, and on existing and new spallation sources) and also in the use of x-ray scattering methods at synchrotron sites and new laboratory sources. Our base of data here is rather limited but could easily be expanded. We will be able to track expansion of the use of our methods to these new instruments/sources from the Science Citation Index to give a clear measure of success.

Higher work throughput and outputs for researchers and research teams 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 we can turn an estimated typical one-day latency of data access down to five minutes.

Improved software and tools. In Benefits Case Study 2, the benefits for code development come through having a wide range of available examples. For example, we need ready access to data for magnetic materials, non-crystalline materials, and materials containing

[12] http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2-ResearchActivityLifecycleModel-110407.pdf

[13] See http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2_BenefitUseCases_final.pdf

molecules, for both neutron and x-ray data. The two markers of success are a) as we develop new functionality we can turn around a suite of test data; b) that as new types of systems demand new functionality, we have the ability to add new data sets to our 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.

3.3 CONCLUSIONS

The I2S2 Benefits Case Studies have been able to illustrate a range of positive benefits that have or would accrue in future from implementation of I2S2. The researcher and service perceptions of benefits can be different but are complementary and together provide a strong argument for further development of I2S2.

The identified benefits can be divided into two major areas:

- Improved research effectiveness including faster information/data access [reduced latency], support for data publication & citation, new research, data training materials, and improved research tools;
- Research support efficiencies including indirect cost savings from increased service productivity or data re-use.

We note the full impact of many benefits cannot always be measured within the timeframe of short projects such as I2S2. Where appropriate we have established benchmarks against which future progress can be measured.

We also note the importance of the end-to-end integration across the community provided by I2S2. To maintain this, the implications for research funder' policy and planning and for community cohesiveness and co-ordination need to be reviewed and actioned on a regular basis.

4 OPTIONS

This section sets out options for sustaining the key deliverables and achievements from the I2S2 project.

4.1 DO NOTHING

Do nothing would require no new capital investment to implement I2S2. However research efficiencies we have identified that would arise from implementation would not occur. Similarly considerable impacts on the effectiveness of research would be lost. There is strong support from the user community for implementation and we believe the capital investment required (see section 5) would be justified by the benefits accrued.

4.2 IMPLEMENTATION AT STFC, DIAMOND AND ISIS

Work is ongoing at STFC to exploit ideas and tools arising from the I2S2 project within STFC facilities and also in facilities internationally. In particular, the following lines of work are underway.

- The ICAT tool set is under continued development to support the data management needs of ISIS and DLS. For example, the ICAT tools have recently been integrated with the Mantid framework[14]. Mantid provides a platform that supports high-performance computing on neutron and muon data providing a set of common services, algorithms and data objects. ICAT can be used with Mantid to search and access raw data sets. Mantid allows the data analysis and stores the resulting data set with a catalogue record in ICAT linked to the dataset, a mechanism explored in I2S2. ICAT is now under review to determine enhancements for ICAT release 4.0. It is anticipated that the enhancements to support data analysis explored in I2S2 will form a major role in this system.
- STFC are lead partner in a European cross-facility initiative PaN-Data[15]. This is a programme to share and co-develop data infrastructure across European photon and neutron facilities, with the aim of integrating their services. The current PaN-Data project, the Strategic Working Group, is developing standards, policies and roadmaps for future integration. Brian Matthews leads the Integration workpackage and is using the results of I2S2 as a basis for integration of research outputs. This will provide the roadmap for the next phase of the programme, the Open Data Infrastructure project (PaNdata ODI) beginning in the autumn of 2011; this project has major workpackages on provenance and preservation and again will exploit the work of the I2S2 project, including software development.
- STFC is also exploiting the outputs of I2S2 in the context of the European project SCAPE, looking at preservation architectures. This is considering the preservation of scientific lifecycles, with a major use case involving the ISIS facility.

[14] http://www.mantidproject.org/Main_Page

[15] http://www.pan-data.eu/

4.3 Implementation at NCS

A pilot implementation is being built at the NCS. The NCS is a "Mid-Range Facility" funded under the eponymous RCUK/EPSRC programme – this funding scheme runs until 2014 and preliminary indications are that a further 5 years of funding (until 2019) would be made available to build on the significant initial investment from the current first phase. As part of this new programme, the NCS is committed to updating its data and information management systems – the intention is to replace and combine the existing systems with a unified framework, underpinned by the I2S2 information model that supports all aspects of the facility operation. The facility operation covers all of the I2S2 research activity lifecycle, from users making an application to use the facility through to dissemination and reuse of results data and crucially supports both user facing activities and interaction and also the service operation and administration. Developer effort from the I2S2 project has initiated the build of this new management system and the resulting pilot implementation will be adopted by the NCS and turned into an operational system. The NCS has both developer effort to enable delivering this service and also systems administration to ensure it is sustained for as long as it is funded.

4.4 Implementation of Impact and Benefits Analysis Tools

We have been successful in obtaining grant funds of £55,420 from the JISC for the "Digital Preservation Benefit Analysis Tools" project which will run from 1st February to 31 July 2011. The project aims to test, review and promote combined use of the Keeping Research Data Safe (KRDS) Benefits Framework and the Value Chain and Impact Analysis tool first applied in the I2S2 project for assessing the benefits and impact of digital preservation of research data. We are extending their utility to and adoption within the JISC community by providing user review and guidance for the tools and creating an integrated toolset. The project consortium consists of a mix of user institutions, projects, and disciplinary data services committed to the testing and exploitation of these tools and the lead partners in their original creation.

The project partners are UKOLN and the Digital Curation Centre at the University of Bath, the Centre for Health Informatics and Multi-professional Education (CHIME) at University College London, the UK Data Archive (University of Essex), the Archaeology Data Service (University of York), OCLC Research, and Charles Beagrie Limited. The tools will be user tested, documented, made freely available, promoted by a range of services, and have value-added support via consultancy if required.

4.5 Maintaining, Extending and Training the I2S2 Community

We have been successful in raising awareness of the issues associated with developing an integrated service approach to managing research data within the I2S2 partner community (largely focussed around chemistry), through project development activities, workshops and conference papers. However, the I2S2 community is only one small part of the much wider structural science community, and a co-ordinated programme of awareness-raising, training, professional development and networking across all the structural science domains, is required to fully realise the I2S2 vision. These might include physics, mineralogy, earth sciences and some aspects of bio-engineering. Extending the I2S2 approach beyond the chemistry domain is a priority.

The EPSRC and STFC research councils are the main UK funding agencies in the structural sciences and we believe that the outcomes from the I2S2 Project have implications for both parties, at both a strategic and policy level, and also at a grassroots research practice level. Any further work in this area should address both aspects. For new-entrant researchers, the EPSRC have funded 50 doctoral training centres (DTC) since 2009 and we believe the DTCs should form important nodes in any structural science data community network. In addition, the Vitae organisation is seen as a key partner in supporting the community dissemination and training programme for doctoral researchers and other research staff.

Key elements in this proposed community programme are:

- Strategic planning advocacy which targets key policy makers and funders including EPSRC and STFC, and for senior PIs within selected institutions
- Coalface / lab-based advocacy which targets new-entrant researchers
- Development of supporting advocacy, training and dissemination materials, including planning guidelines, exemplar mini case studies, toolset guides
- Work with local DTCs to develop and test training resources
- Embed research data management training in the postgraduate curriculum, through DTC workshops, lab visits and surgeries
- Emergence of a self-sustaining structural science data community network facilitated by face-to-face events and appropriate social network tools.

4.6 Conclusions

We have concluded that there is merit in adopting an integrated approach which caters for all scales of science and there is a strong case for seeking to implement and further develop the outcomes from I2S2. Ideally, it would be beneficial to implement a second phase of the work begun in I2S2 to fully implement the pilot infrastructure components developed to date and embed them within the Structural Science community in a coordinated manner. This would however require project funding on a timescale of 2-3 years to achieve. Given the uncertainty of appropriating the necessary funding, we have considered alternative ways of sustaining, maintaining and further developing specific aspects of the work begun in I2S2 in the medium-term through projects and initiatives that are currently on the immediate horizon for I2S2 partner organisations.

5. Costs and Timescales

5.1 Implementation at STFC, Diamond and ISIS

As noted in section 4.2, STFC intends to leverage the ideas and continue development of I2S2 tools through further projects and initiatives; this will involve no additional cost to funding agencies. It will also allow independent support for and evolution of the relevant work over a time-scale of 1-3 years.

5.2 Implementation at NCS

The NCS pilot implementation will be adopted and used as an operational service, which will be sustained for the lifetime of the facility **at no extra cost to funding agencies** (see 4.3 for explanation).

The NCS was the first facility to be funded under the new Mid-Range Facilities programme and in many ways is acting as an example of best practice for the following services – this is especially so in the area of data management, due to a long track record of NCS involvement in such activities (eBank-UK, eCrystals, R4L, oreChem, I2S2, IDMB). Once all the services in the programme have been appointed, RCUK will hold a best practice workshop (Dec 2011) and the NCS has been invited to present the I2S2 system as the exemplar of data management, user interaction and service operation.

There is therefore a process in place for disseminating the outputs of I2S2 and advocating their use to the services in the EPSRC/RCUK Mid-Range Facilities programme within 6 months of the end of the project. Most of the underlying work for implementing I2S2 in other services has been performed through the derivation of the I2S2 information model. However, each facility is different in that it provides different services to different user bases (both scientifically and in terms of scale) and therefore there is no 'one size fits all' solution. Building on the I2S2 foundations it would take 1FTE developer 2-3 months to implement a data management solution for a moderately sized service.

5.3 Implementation of Impact and Benefits Analysis Tools

As noted in section 4.4 we have already been successful in obtaining funds and building partnerships to implement benefit and impact tools and leverage work in I2S2 in this area. We anticipate this will provide the foundation needed for further adoption and testing within the community and the potential to build on emerging work on defining metrics if future funding opportunities arise.

5.4 Maintaining, Extending and Training the I2S2 Community

A timeframe of three years would be an ideal period for full implementation and evaluation of the proposed programme, since the programme would need to be trialled over a 12 month period, evaluated, any modifications made and then sustained in the revised format, either by institutions through embedding within doctoral training programmes, or by the research funding bodies through their professional development funds.

Total estimated costs for the above are £60K, which include programme development, delivery, evaluation, revisions and embedding.

5.5 Maintaining The Information Model

The information model is a work in progress; the current version is a snapshot of its development as it stands, and it will be continued to be worked on in current projects such as PaN-Data and the UMF Smart Research Framework, which will sustain and promote the model for the next 2-3 years, and build tools which will use and develop it further. Ongoing support of the model will be dependent on its value to the community; if it has wider take up we will develop a sustainable model. We would anticipate that we would seek to support it within an open source community effort, such as the SPAR ontologies. For example, the

model already has a space within the ICAT open source project; we would use this open space to promote and develop the model with interested parties.