Project Document Cover Sheet

Project Information			
Project Acronym	I2S2		
Project Title	Infrastructure for Integration in Structural Sciences		
Start Date	1st Oct 2009	**End Date**	31st July 2011
Lead Institution	Universities of Bath and Southampton		
Project Director	Liz Lyon (UKOLN)		
Project Manager & contact details	Manjula Patel, UKOLN, University of Bath, Tel: 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 Project Plan		
Reporting Period	*for progress reports only*		
Author(s) & project role	Manjula Patel (Project Manager)		
Date	05/11/2009	**Filename**	I2S2ProjectPlan-091105
URL	*if document is posted on project web site*		
Access	X Project and JISC internal		□ General dissemination

Document History		
Version	**Date**	**Comments**
0.1	13/10/2009	Initial version
0.2	28/10/2009	Amendments following kick-off meeting
0.3	30-31/10/2009	Version for comments and input from partners
0.4	02/11/2009	Added in budget
0.5	05/11/2009	Amendment according to comments from partners

JISC Project Plan

Overview of Project

1. Background

The I2S2 project aims to understand and identify the requirements for a data-driven research infrastructure in the Structural Sciences. The project also seeks to show how effective cross-institutional research data management can increase efficiency in the use of multi-million pound centralised facilities such as the DIAMOND Light Source (DLS) and ISIS facilities at STFC.

'Structural Science**'** in its widest sense transcends, encompasses and underpins Physical, Earth, Materials and Life Sciences as shown in Figure 1.

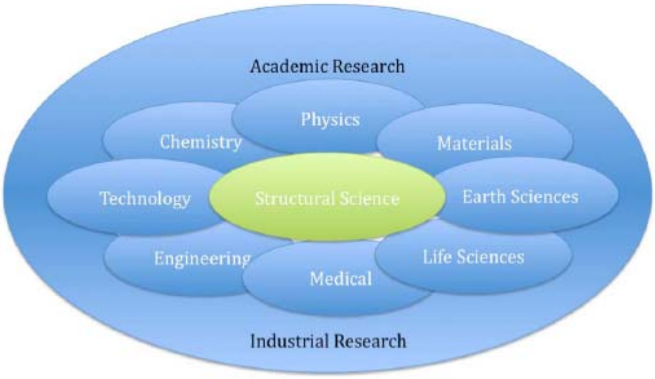

Figure 1: Structural Science Facets

The work will focus on the exemplar domain of Chemistry, but with a view towards inter-disciplinary application. Current practice in Chemistry is such that:

- Research teams capture, manage, discuss and disseminate their data in relative isolation with highly fragmented data infrastructures and poorly integrated software applications;
- Strict adherence to the conventional systems of publication leads to insufficient information on results (especially for those outside an established core of researchers) and to irreproducible experiments;
- The processes for recognition lead to a lack of inclination and incentive to share or make all the supporting information for a study publicly available;
- A low awareness of data curation and preservation issues leads to data loss and reduced productivity.

This culture of fragmentation and conventionality means that there are insufficient opportunities for innovative multi-disciplinary research and discovery. The chemistry community is ideally placed to benefit from a step change in working practice resulting from an integrated data management infrastructure. Similar behaviour is observed across the Physical and Earth Sciences in stark contrast to the Bio-sciences, which have embraced 'bio-informatics' and provide an indication of the potential for physical sciences.

Structural Science is generally concerned with three aspects of research practice:

a) Performing experiments: either in the local institutional laboratory or at large central facilities.
b) Performing simulations: commonly executed using high-throughput instrumentation with many parallel instances (e.g. based at the NCS or local computing facilities), high-performance computing (national or university facilities), or private computing (laptop or desktop).
c) Analyses of both experiment and simulation results, including integration of data from different experiments and/or simulations to produce a combined analysis to determine different features, typically using resources within the scientist's personal data workspace.

Departmental research groups may have localised data management practices which result in major differences in the way data are collected and preserved across the range of institutions that comprise a research community. Moreover, large centralised facilities have a responsibility to provide a data management infrastructure for their users and have spent considerable effort designing and implementing such systems. The outcome is that each central facility has its own, very insular, approaches to data management, which result in vast 'data silos'.

2. Aims and Objectives

In broad terms, the project will provide the community with pilot data management infrastructure solutions which bridge discipline, laboratory and institutional boundaries.

More specifically, the project will develop and test transformative data practices across the whole research data lifecycle in structural science by using the domain of Chemistry to:

- Develop a framework for data management, deployable across the full range of Structural Science;
- Explore a range of data acquisition techniques at different scales (complexity, volume, definition);
- Promote common and easy access to data, experimental and computational resources through the infrastructure, so that transaction costs can be minimised;
- Advocate recognition in the community for sharing data to encourage its reuse;
- Enable discovery of results in related disciplines via common terms;
- Facilitate access to data underpinning publications with higher levels of verification, resulting in higher quality research;
- Enhance rapid communication across the community;
- Support long-term preservation assuring future discovery of results.

3. Overall Approach

The I2S2 project aims to address three complementary infrastructure "axes" as shown in Figure 2:

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

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

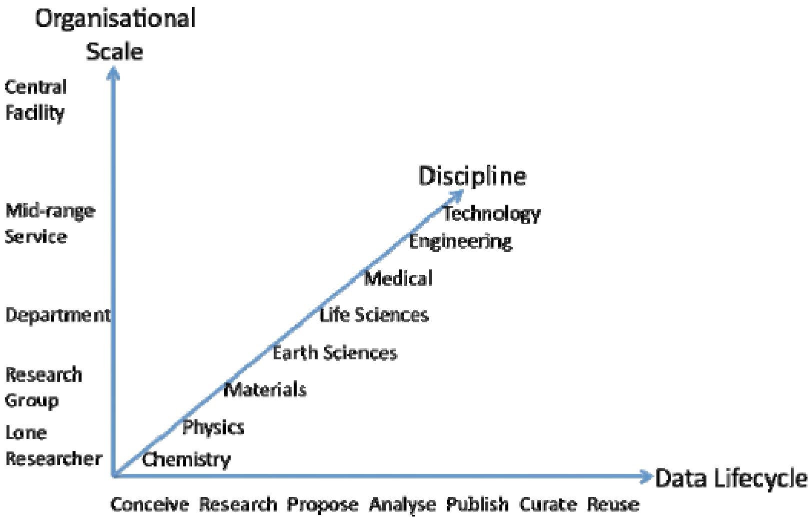

Figure 2**:** Three Dimensions of Research Data Management Infrastructure

The DLS is a third generation synchrotron radiation source operated by Diamond Light Source Ltd, a company co-owned by STFC. It is a state of the art scientific facility that is enabling scientists to investigate the structure of matter, such as biological tissues, polymers and catalysts, at the atomic and molecular level.

ISIS is a world-leading pulsed neutron and muon source operated by the STFC at Rutherford Appleton Laboratory. The facility enables scientists to probe the microscopic structure and dynamics of matter. ISIS is used by over 2000 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.

The project will develop two use cases that explore the perspectives of “scale and complexity” and “research discipline” throughout the data lifecycle. This will be achieved through implementation of two pilots based on the use cases, which will examine the business processes of research, and highlight the benefits of an integrated approach. Both pilots will address the traversal of administrative boundaries between institutions to national facilities in addition to issues of scale (local laboratory to national facilities, DLS synchrotron and ISIS respectively). Pilot 1 will be a case study involving workflows between the NCS (Soton) and DLS, whilst Pilot 2 will apply the approach to Earth Sciences (Cambridge and ISIS) and demonstrate the benefit to scientific disciplines other than Chemistry.

Key to the achievement of the project’s objectives will be the development of an abstract information model for structural science research as a fundamental infrastructural component driving distributed data management implementations (database schema, registries, APIs, Laboratory Information Management Systems (LIMS), exchange formats etc.) and effective inter-disciplinary data sharing.

STFC e-Science has been working with both ISIS and DLS to introduce data management infrastructure, developing the ICAT suite of tools and the CSMD model (see section 8 on standards) to organise data derived in investigations using the facilities. While the core suite is the same (both now using ICAT version 3.3), its deployment has been tailored to the specific needs of each facility, so while the pilots will need to be specific to each facility, they should remain interoperable and comparable through the use of the core ICAT and metadata components.

In addition to the above, three thematic strands will be followed throughout the study:

a) **Scholarly issues** (Intellectual/Scientific elements including data longevity, data provenance, data mining, knowledge extraction, innovative use)
b) **Economic issues** (costs, value-for-money, investment returns, financial sustainability)
c) **Societal issues** (research practice, cultural change, drivers and barriers).

4. Project Outputs

In addition to the deliverables list below, the project will also provide the community with pilot data management infrastructure solutions as well as providing insights into their implementation and use.

WP	Description	Deliverables	Month	Lead + partners
1	Foundational Principles	D1.1 Requirements Report D1.2 Two Use Cases	1-4	UKOLN DCC
2	Costing Baseline	D2.1 Extended Cost Model D2.2 Cost Analysis Phase 1	3-4	Soton, Beagrie
3	Harmonisation & Implementation	D3.1 Integrated Information Model D3.2 Implementation Plan D3.2 Two Pilot Implementations	4-15	STFC + Camb/Soton
4	Benefits Realisation	D4.1 Cost Analysis Phase 2 D4.2 Benefits Report and Business Model	15-16	Soton, Beagrie
5	Dissemination and Outreach	D5.1 Advocacy and Training materials D5.2 Two Workshops	1-18	UKOLN DCC+ All
6	Project Management, Evaluation & Sustainability	D6.1 Project Plan D6.2 Final Report D6.3 Sustainability Recommendations	1-18	UKOLN DCC + All

5. Project Outcomes

The practical outcomes of the project have the potential to greatly streamline the working processes of structural scientists in many domains and at many locations.

The relationship between HE institutions carrying out research in local laboratories and associated research executed at central facilities such as synchrotrons is not well-understood. There are opportunities for I2S2 to contribute towards a better understanding of how the various stakeholders can work more effectively together both at policy and practitioner levels, to enhance data integration, build more productive partnerships, generate cost savings, improve research productivity and ultimately do better science. The development and embedding of economically efficient data management processes underpinned by a robust data infrastructure is a major first step in achieving these aspirations. In the longer term, there are exciting possibilities in extending the relatively modest work proposed in I2S2 to the more challenging volumes of data which will be processed through DIAMOND as more beam lines become active.

6. Stakeholder Analysis

Stakeholder	Interest / stake	Importance
JISC	Funding	High
Other JISC Projects	I2S2 deliverables may be of interest	Medium
Practitioners in university departments at Cambridge and Southampton	Interest in data management infrastructure solutions and best practice	High
Users of Central Facilities	Interest in data management infrastructure solutions, efficiency gains and best practice	High
Users of NCS	Interest in data management infrastructure solutions, efficiency gains and best practice	High
Researchers in Structural Sciences	Interest in data management infrastructure solutions and best practice	High

7. Risk Analysis

Risk	Probability (1-5)	Severity (1-5)	Score (P x S)	Action to Prevent/Manage Risk
Staffing	1	4	4	Existing staff will be drawn upon where possible in case of loss of staff or absence due to sickness.
Organisational	1	3	3	Project management team already in place. Consortium agreement and Virtual Advisory Board to be established within first 3 months of project.
Technical	2	3	6	Technical work is spread over several partners. Multiple technology paths will be identified.
External suppliers	1	1	1	The project is not dependent on external suppliers
Legal	2	4	8	Intellectual property issues will be referred to JISC, DCC and institutional legal experts

8. Standards

Name of standard or specification	Version	Notes
DCC Data Management Plan checklist	2.0	
DCC Data Audit Framework		
DCC Curation Lifecycle Model		
PREMIS	2.0	Preservation metadata standard
Keeping Research Data Safe Model (KRDS)	2.0	
STFC Core Scientific Metadata Model (CSMD)	3.0	Metdata specification used within STFC facilities. Proposed as a reference model for I2S2
ICAT		Open source software released by STFC
Chemical Mark-up Language (CML)		
UML, XML, RDF		
CIF, checkCIF, InChi		

9. Technical Development

The project will adopt open standards as far as possible. Any software developed exclusively within the project will be released under an appropriate open source licence. I2S2 is not a primarily software development project, but it may be necessary to implement software solutions to support integration and interoperability.

10. Intellectual Property Rights

The project will comply with the JISC Funding Agreement. It is expected that whilst most outputs will be openly available (with Creative Commons licenses where appropriate) on the I2S2 website, any intellectual property from resulting research outputs, will be subject to the Copyright, Designs and Patents Act 1988. We are aware of IPR issues which may arise from cross-sectoral collaborations and will seek expert advice from institutional and DCC legal experts.

Project Resources

11. Project Partners

The consortium agreement will be signed by 31st December 2009.

UKOLN / Digital Curation Centre, University of Bath

Liz Lyon is Director of UKOLN. She led the eBank UK project, is Associate Director (Community Development) of the UK Digital Curation Centre (DCC), and authored the *Dealing with Data* Report. She has a doctorate in cellular biochemistry. **Manjula Patel** has worked on the eCrystals Project looking at the curation and preservation of crystallography research data and on the EPSRC KIM project investigating engineering data curation issues.

Liz will lead the contribution from UKOLN/DCC; Manjula will act as Project Manager and Research Officer on the project.

University of Southampton
Simon Coles directs the UK National Crystallography Service and has been PI on numerous e-research and e-learning projects on data driven science since starting work in the field as a Co-Investigator on the EPSRC eScience pilot project, CombeChem. UoS effort will also draw on the expertise of **Professor Jeremy Frey** (PI, CombeChem) who is a world leader in developing innovative approaches for research in the digital environment. Simon will lead the Soton contribution.

University of Cambridge
Martin Dove led eScience/Informatics projects including eMinerals, MaterialsGrid, National Institute for Environmental eScience (just ended). Martin is a long-standing and significant user of ISIS. His research group is in Earth Sciences. **Peter Murray-Rust** and **Jim Downing** work in the Unilever Centre and lead the CLARION project using Electronic Laboratory Notebooks. Martin and Peter will lead the contribution from Cambridge.

Science & Technology Facilities Council
Brian Matthews is leader of the Scientific Applications Group in the e-Science Centre and has a major role in development projects to support facilities science, including developing the CSMD metadata model for science data, federated authentication across STFC, and the ePubs archive. Brian will lead the STFC contribution.

Charles Beagrie Ltd.
Neil Beagrie is founding director of Charles Beagrie Ltd. and a leading expert on digital preservation and curation with an international reputation and many international clients such as the EC and Library of Congress. He has been project leader for the JISC Keeping Research Data Safe study. Neil will lead the contribution from Charles Beagrie.

12. Project Management

UKOLN/DCC will provide project management capability and day-to-day operational oversight of the work. Project start-up will be informed by an initial face to face (F2F) meeting of all project personnel at month 1 to establish the Project Plan, with further full F2F meetings at 9 & 15 months. This will facilitate an evaluation of the requirements and final design/refinement of the use cases; a mid point health check (particularly on progress of WP3); critical evaluation of the cost-benefit analysis; review of the demonstrator implementation and preparation for report writing. Monthly project telephone conferences together with bilateral partner meetings will provide practical management between full F2F meetings.

A ‘virtual’ Advisory Board will be convened for steering the project and providing additional links with the community. The project team will work pro-actively with JISC and Simon Hodson will be invited to join the Advisory Board. Reports will be provided to the JISC as required.

Project Team & Contact Details

Elizabeth Lyon (PI), University of Bath, UKOLN, tel:01225 386547, fax:01225 386838, E.J.Lyon@ukoln.ac.uk

Simon Coles (PI), University of Southampton, EPSRC National Crystallography Service, tel:023 8059 4479, fax:023 8059 2865, s.j.coles@soton.ac.uk

Manjula Patel (Project Manager (0.15FTE) and Research Officer (0.35FTE)), University of Bath, UKOLN, tel:01225 386254, fax:01225 386838, m.patel@ukoln.ac.uk

Neil Beagrie (CI), Charles Beagrie Ltd., tel:0709 204 8179, fax: 0709 204 8179, neil@beagrie.com

Brian Matthews (CI), STFC, Scientific Applications Group, tel:01235 446648, fax:01235 445945, brian.matthews@stfc.ac.uk

Martin Dove (CI), University of Cambridge, Dept. Earth Sciences, tel:01223 333482, fax: , mtd10@cam.ac.uk

Juan Bicarregui (CI), STFC, e-Science Applications Support Division, tel:01235 445710, fax:01235 445945, juan.bicarregui@stfc.ac.uk

Peter Murray-Rust (CI), University of Cambridge, Dept. Chemistry, tel:01223 763069, fax: pm286@cam.ac.uk

Training in project management may be beneficial to the Project Manager. It may be possible to attend an appropriate course run by the JISC InfoNet.

13. Programme Support

The project will benefit from participation in Programme level workshops enabling knowledge and experiences to be shared amongst related projects. We will also draw on the expertise and support of the Programme Support Projects.

14. Budget

See Appendix A.

Detailed Project Planning

15. Workpackages

See Appendix B.

16. Evaluation Plan

Evaluative processes have been built into the project through a cost-benefit analysis in WPs 2 and 4. This will result in a comparison of costs before and after the implementation of the pilots which will demonstrate the efficiency of the framework being implemented. A self-evaluation of the whole project will be carried out towards the end of the work which will also include commentary collected from the wider community.

Timing	Factor to Evaluate	Questions to Address	Method(s)	Measure of Success
WP2, WP4	Cost/benefit	Whether implementation of the data management infrastructure has	Cost/Benefit analysis	“Before” and “After” Activity Based Costings; Identification of

		resulted in net gains		quantitative and qualitative benefits
End of project	Overall impact of project	Are stakeholders on board? What benefits are there for stakeholders? Have objectives been met? Have outcomes been achieved? What are the key findings? What impact did the project have?	Self-evaluation and Community commentary	

17. Quality Plan

Output	**Reports**				
Timing	**Quality criteria**	**QA method(s)**	**Evidence of compliance**	**Quality responsibilities**	**Quality tools (if applicable)**
Project duration	Fitness for purpose	Review by project partners	Usefulness and impact on other stages of project	Report author(s) + Project partners	Spelling and grammar checking tools

Output	**WP3: Integrated Information Model**				
Timing	**Quality criteria**	**QA method(s)**	**Evidence of compliance**	**Quality responsibilities**	**Quality tools (if applicable)**
WP3	Fitness for purpose	Review by project partners; Testing	Ability to support pilot implementations	WP3	

Output	**WP5: Advocacy and Training materials**				
Timing	**Quality criteria**	**QA method(s)**	**Evidence of compliance**	**Quality responsibilities**	**Quality tools (if applicable)**
WP5	Fitness for purpose	Feedback from users at NCS and STFC	Impact on stakeholder community	Authors and trainers	
	Usability	Peer review	Take up by	Authors and	

			trainers	trainers	

18. Dissemination Plan

Timing	Dissemination Activity	Audience	Purpose	Key Message
Months 2-4	Requirements gathering	Scientists in structural science	Engagement with the project	
Months 1-18	Articles, presentations, and keynotes at various meetings and conferences e.g. DCC Conference, UK eScience AHM, IEEE eScience, CNI Taskforce, eResearch Australasia.	Scientific and research communities in eScience, digital preservation etc.	Advocacy and awareness raising	
Months 1-18	Advocacy and tutorial materials	Scientists in structural science	Training and promotion of best practice	
Months 16-18	Disciplinary community workshop	Scientists in structural science	Advocacy and awareness raising	
Months 16-18	Workshop in association with Research Data Management Forum	Scientific and research communities in structural science, eScience, LIS, digital preservation etc.	Advocacy and awareness raising	

19. Exit and Sustainability Plans

Project Outputs	Action for Take-up & Embedding	Action for Exit
Project website		To be maintained by UKOLN for 3 years after project ends
Reports	Effective dissemination and promotion throughout stakeholder communities	Ensure final versions readily available on project website
Advocacy and tutorial materials	Effective dissemination and promotion throughout stakeholder communities	Ensure final versions readily available on project website

Project Outputs	Why Sustainable	Scenarios for Taking Forward	Issues to Address
Data management	We expect the	Incorporation into	Operational aspects

infrastructure will continue to be used by NCS (between Soton and DLS)	Pilots to demonstrate advantages relating to scale, complexity and inter-disciplinary use.	processes and procedures at NCS and DIAMOND	
Data infrastructure framework	May be useful to stakeholder communities	Continued embedding and development by the community	
Information Model	Underpins the data management infrastructure	Maintenance and development by STFC, DCC or using a community social model	
Cost/Benefit Model	May be useful to stakeholder communities and JISC	Maintenance and development by JISC or Charles Beagrie Ltd	

Appendixes

Appendix A. Project Budget

JISC

Appendix B. Work Packages

JISC WORK PACKAGE

WORKPACKAGES		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	Month	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M
1: Foundational Principles		■	■	■	■														
2: Costing Baseline				■	■														
3: Harmonisation & Implementation					■	■	■	■	■	■	■	■	■	■	■	■			
4: Benefits Realisation																■	■		
5: Dissemination and Outreach		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
6: Project Management, Evaluation & Sustainability		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

Project start date: 1st October 2009

Project completion date: 31st March 2011

Duration: 18 months

Work Package and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
YEARS 1 and 2					
WORKPACKAGE 1: Foundational Principles *Objective*:			**D1.1 Requirements Report** **D1.2 Two Use Cases**		**UKOLN/DCC + ALL**
1. Requirements Analysis	1/11/2009	1/12/2009	Synthesis of requirements analysis		MP/NB
2. Gap Analysis	1/12/2009	31/12/2009	Gap analysis report		MP
3. Immersive Studies	1/11/2009	31/12/2009	Immersive case studies report		MP
4. Use cases	1/1/2010	31/1/2010	Two use cases		MP/SC/MD/PMR
WORKPACKAGE 2: Costing Baseline **Objective:**					**SOTON**
1. Extending the Keeping Research Data Safe (KRDS) Model	1/11/2009	31/12/2009	**D2.1 Extended Cost Model**		NB/SC
2. Metrics and assigning costs	1/11/2009	31/03/2010	**D2.2 Cost Analysis Phase 1**		NB/SC + ALL
WORKPACKAGE 3: Harmonisation of Infrastructure Models & Implementation **Objective:**					**STFC + ALL**
1. Integrated Information Model a. Develop, Conceptualise and Creation planning b. Data Curation and Preservation	1/1/2010	1/4/2010	**D3.1 Integrated Information Model**		BM/SC/MP

planning c. Data access and data flows					
2. Development of Pilots	1/4/2010	31/12/2010	**D3.2 Implementation Plan** **D3.3a Pilot One: Scale and Complexity (Chemistry)** **D3.3b Pilot Two: Inter-disciplinarity (Earth Science)**		BM/SC BM/SC/PMR BM/SC/MD
WORKPACKAGE 4: Benefits Realisation **Objective**:					**SOTON**
1. Post-Pilot Implementation Benchmark Costs	1/12/2010	31/1/2011	**D4.1 Cost Analysis Phase 2**		NB/SC
2. Benefits Analysis	1/12/2010	29/02/2011	**D4.2 Benefits report and business model**		NB/SC + ALL
WORKPACKAGE 5: Dissemination and Outreach **Objective**:					**UKOLN/DCC + ALL**
1. Outreach and Community Engagement Programme	1/10/2009	31/3/2011	**D5.1 Advocacy and Training materials**		LL/SC
2. Workshops at Rutherford Appleton Laboratory	1/1/2011	31/3/2011	**D5.2 Two Workshops** 1. Disciplinary Community (Structural Science conferences - BCA Spring Meeting (Warwick, April 2010 and Keele, April 2011)) 2. With RDMF		SC/BM LL/BM

WORKPACKAGE 6: Project Management, Evaluation and Sustainability **Objective**:					**UKOLN/DCC + ALL**
1. Project management	1/10/2009	31/3/2011	**D6.1 Project Plan** **D6.2 Final Report** **Progress Reports etc.**		MP
2. Evaluation	1/1/2011	1/3/2011	**Evaluation report**		LL/MP
3. Sustainability	1/1/2011	31/3/2011	**Sustainability Recommendations**		LL/MP

Members of Project Team:

LL = Liz Lyon	JB = Juan Bicarregui
SC = Simon Coles	MD = Martin Dove
NB = Neil Beagrie	PMR = Peter Murray-Rust
BM = Brian Matthews	MP = Manjula Patel