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**	31st July 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 Project Plan		
Reporting Period	*N/A*		
Author(s) & project role	Manjula Patel (Project Manager)		
Date	7th July 2010	**Filename**	I2S2-WP1-D1.1-RR-Final-100707.doc
URL			
Access	X Project and JISC internal		☐ General dissemination

Document History		
Version	**Date**	**Comments**
0.9	7th July 2010	Final version submitted to JISC

Infrastructure for Integration in Structural Sciences

D1.1 Requirements Report

Work Package 1

November 2009 – June 2010

JISC I2S2 Project

Document Details

Author:	Manjula Patel (UKOLN & DCC)
Date:	7th July 2010
Version:	0.9
File Name:	I2S2-WP1-D1.1-RR-Final-100707.doc
Notes:	Final version for submission to JISC

Acknowledgements

The Infrastructure for Integration in Structural Sciences (I2S2) Project is funded by the UK's Joint Information Systems Committee (JISC); the Programme Manager is Simon Hodson. The I2S2 project team comprises:

- Liz Lyon (UKOLN, University of Bath & Digital Curation Centre)
- Manjula Patel (UKOLN, University of Bath & Digital Curation Centre)
- Simon Coles (EPSRC National Crystallography Centre, University of Southampton)
- Martin Dove (Earth Sciences, University of Cambridge)
- Peter Murray-Rust (Chemistry, University of Cambridge)
- Brian Matthews (Science & Technology Facilities Council)
- Erica Yang (Science & Technology Facilities Council)
- Juan Bicarregui (Science & Technology Facilities Council)
- Neil Beagrie (Charles Beagrie Ltd.)

Executive Summary

The *Infrastructure for Integration in Structural Sciences (I2S2) Project* is funded under the Research Data Management Infrastructure strand of the JISC's Managing Research Data Programme, with a duration of 18 months (Oct 2009 to March 2011).

One of the main aims of the project is to investigate the research and data management infrastructure needs of researchers in the Structural Sciences (incorporating a number of disciplines including Chemistry, Physics, Materials, Earth, Life, Medical, Engineering, and Technology). We define research infrastructure to encompass physical, informational and human resources essential for researchers to undertake high-quality research, including: tools, instrumentation, computer systems and platforms, software, communication networks, technical support (both human and automated); documentation and metadata.

It is important to realise that the life expectancy of scientific data has increased over the years as more and more scientific research becomes derivative in nature, dependent on data generated, managed and made widely accessible to third parties. However, effective reuse and repurposing of data requires much more information than the dataset alone. Trust and a thorough understanding of the data is a precursor to its reuse and this in turn necessitates transparency and access to considerable contextual information regarding how the data was generated, processed, analysed and managed. Consequently, within the I2S2 project, research data is considered to be not only the raw images and numerical datasets that are generated and collected from scientific experiments, but also the broader categories of information that are associated with such data. Various types of related data have been identified through the development of an idealised activity lifecycle model for scientific research.

In order to facilitate the investigation of scale and complexity, inter-disciplinary and data lifecycle issues, the project incorporates partners currently working at differing scales of science, as well as in differing disciplines: University of Cambridge (Earth Sciences) represents a lone scholar (or small team) scenario; University of Cambridge (Chemistry) may be considered at a large research group or department level; the EPSRC National Crystallography Service (NCS) is an example of a mid-range service facility; whilst the DLS and ISIS at the Science & Technology Facilities Council (STFC) represent large-scale central facilities. The project is focusing on the domain of Chemistry, but with a view towards inter-disciplinary application by understanding localised data management practices in research institutions as well as the data management infrastructures in large centralised facilities.

This document (deliverable D1.1, Requirements Report) presents the results of the work undertaken to identify requirements for the I2S2 project. We report details of the activities largely carried out between November 2009 and April 2010 to progress requirements capture and analysis. In addition, we describe the methodologies used to identify the requirements, the initial findings and an analysis of the results. The results of this analysis will feed into the development and implementation of two pilot infrastructures (deliverable D3.2) which will be based on two use cases (deliverable D1.2) and an integrated information model (deliverable D3.1).

Despite the considerable variation and diversity in requirements between the different scales of science being undertaken, a relatively common thread has become apparent in the form of a need to be able to manage all data as they are generated, collected and processed during the course of scientific research experimentation. We provide an itemised summary of the findings of the requirements gathering process below, additional details and analyses can be found in the main text of the report:

- A robust data management infrastructure which supports each researcher in capturing, storing, managing and working with all the data generated during an experiment.
- Department or research group level data storage, backup and management facilities.

- Internal sharing of research data amongst collaborating scientists, such as between a PhD student and supervisor.
- Sharing of data with third parties.
- Access to research data in the long run so that a researcher (or another team member) can return to and validate the results in the future.
- Capture, management and maintenance of:
    - Metadata and contextual information (including provenance)
    - Control files and parameters
    - Versioning information
    - Processing software
    - Workflow for a particular analysis
    - Derived and results data
    - 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.
- Where crystallography data repositories already exist, there is a requirement to develop them into a robust service incorporating curation and preservation functions.
- There is a real need for IPR, embargo and access control to facilitate the controlled release of scientific research data.
- Valuable information commonly stored in analogue laboratory notebooks is difficult to share and reuse and needs to be stored digitally.
- The potential of data for reuse and repurposing could be maximised if standard data formats and encoding schemes, such as XML and RDF, are widely used.
- Paper and or hybrid record-keeping and resource scheduling systems would benefit from automated processing.
- Use of consistent and persistent identifiers would greatly aid the seamless flow of information between organisations, applications and systems.
- There is a need to streamline administrative functions between organisations, for example through the use of standardised Experiment Risk Assessment forms (ERAs).
- It is clear that the processing pipeline in many scientific experiments tend to be near digital, relying on suites of tools, applications software and very often customised software. There is therefore a need to document, maintain and curate such software.
- The Core Scientific Metadata Model (CSMD) and its implementation in ICAT is a good candidate for further development and extension to take account of the needs of organisations outside of the STFC.

The four broadly defined levels of research science examined in the report (individual researcher, team, and medium-level service to large-scale facility) reveal 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 mange 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.

We conclude that there is merit in adopting an integrated approach which caters for all scales of science (although the granularity and level of integration is an area that needs further investigation). Furthermore, an integrated approach to providing data management infrastructure would enable an efficient exchange and reuse of data across disciplinary boundaries; the aggregation and/or cross-searching of related datasets; and data mining to identify patterns or trends in research and experiment results.

In addition, demands are now surfacing for “Open Methodology”, such that making data alone openly available is insufficient and there are now expectations that the methodologies used in processing and analysing them should also be made readily accessible. The work being undertaken in the I2S2 project (in terms of building a robust data management infrastructure) has the potential to form a foundation on which differing methodologies can be both run and exposed to third parties for easier sharing.

Document Revision History		
Version	**Date**	**Comments**
0.1	18-21st Jan 2010	Scope and initial outline
0.2	27th Jan -10th Feb 2010	Proposed research data lifecycle model V1.0 Methodology section Desk study documents Findings section (desk study; immersive studies; gap analysis)
0.3	1st-11th Mar 2010	Proposed research data lifecycle model V2.0 (modifications for comments resulting from presentation at I2S2 Models Workshop held on 11th Feb 2010) Data management planning tools section More on findings section Inclusion of Cambridge Chemistry following site visit on 4th March 2010
0.4	22-31st Mar 2010	Added new logo Research data lifecycle model V4 More on immersive studies and Gap analysis Included Martin and Erica's workflow report as appendix Requirements Analysis
0.5	1st April 2010	Requirements Synthesis Working draft circulated for partner comment
0.6	22-24th June 2010	Research data lifecycle activity model V5 Acknowledgements; Conclusions
0.7	28-29th June 2010	Executive Summary; Citations Circulation of final draft for partner comment
0.8	6th July 2010	Final version of research activity lifecycle model Amendments following partner feedback
0.9	7th July 2010	Separation out of Appendix due to formatting problems

Contents

1. Introduction

One of the main aims of I2S2 is to identify the requirements for a data-driven research infrastructure by understanding localised data management practices in research institutions as well as the data management infrastructures in large centralised facilities. We define research infrastructure to encompass physical, informational and human resources essential for researchers to undertake high-quality research, including: tools, instrumentation, computer systems and platforms, software, communication networks, technical support (both human and automated); documentation and metadata.

The project proposes an examination of three complementary infrastructure strands, as shown in Figure 1:

> ***Scale and complexity***: from small laboratory bench based science to institutional installations to large scale facilities such as the DIAMOND Light Source (DLS) and ISIS, based at the Science and Technology Facilities Council (STFC) at the Rutherford Appleton Laboratories in Didcot.
>
> ***Inter-disciplinary issues***: research across disciplinary domain boundaries.
>
> ***Data lifecycle***: data flows and data transformations throughout the useful life time of the data.

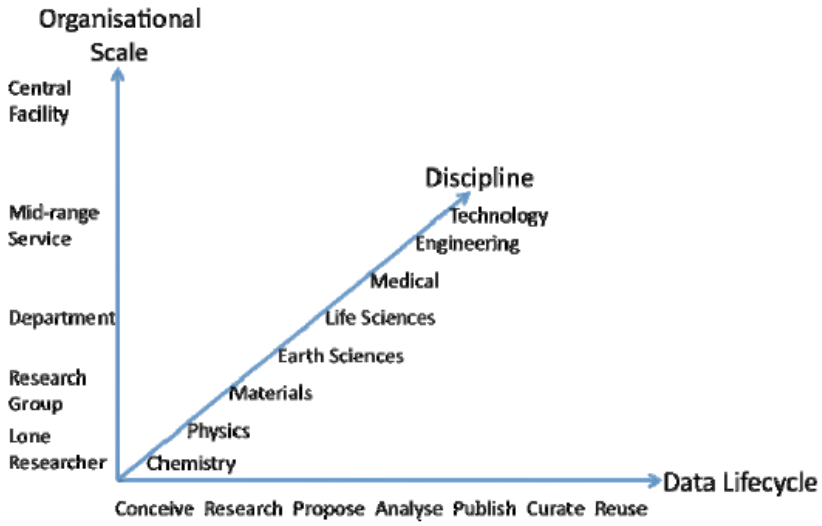

Figure 1: Three complementary axes in I2S2

In order to facilitate the investigation of these three axes, the project incorporates partners currently working at differing scales of science, as well as in differing disciplines: University of Cambridge (Earth Sciences) [1] represents a lone scholar (or small team) scenario; University of Cambridge (Chemistry) may be considered at a large research group or department level [2]; the EPSRC National Crystallography Service (NCS) is an example of a mid-range service facility [3]; whilst the DLS and ISIS at the STFC represent large-scale central facilities [4].

This document (deliverable D1.2, Requirements Report) presents the results of the work package to identify user requirements for the I2S2 project. The main goal is to investigate the research and data management infrastructure needs of researchers in the Structural Sciences (incorporating a number of disciplines including Chemistry, Physics, Materials, Earth, Life, Medical, Engineering, and Technology). As such, this work package aims to elicit the views

of research practitioners in the chemistry and earth science domains, based at the Universities of Southampton and Cambridge. We report details of the activities largely carried out between November 2009 and April 2010 to progress user requirements capture and analysis. In addition, we describe the methodologies used to identify the requirements, the initial findings and an analysis of the results. The results of this analysis will feed into the development and implementation of two pilot infrastructures (deliverable D3.2).

2. Methodology

The Research Data Management Infrastructure (RDMI) strand and its component projects, have a strong emphasis on adopting a user-centred approach to the development of infrastructures for the management of research data. Consequently, the approach of the I2S2 project is to involve specific research practitioners and formulate the user requirements largely from consultation with them, but supported by additional requirements capture processes.

The I2S2 Project Plan [5] outlines several methods for extracting and eliciting requirements for the project including a desk study, immersive studies, a gap analysis, development of use cases and employment of various tools such as the Digital Curation Centre's Data Audit Framework [6] and Data Management Plan checklist [7], as well as the methods promoted in the Keeping Research Data Safe projects [8][9] for performing cost/benefit analyses.

Additionally, the requirements capture process will be driven by two case studies:

Case study 1: Scale and Complexity
This is concerned with traversing the administrative boundaries between institutions and experiment service facilities and will be based around the interactions between a lone worker or research group in their home institution, the EPSRC UK National Crystallography Service or NCS (a mid-scale facility providing experiment data capture for UK academics) and the central facilities synchrotron, the DIAMOND Light Source (DLS) [10]. Modelling these interactions will probe both the cross-institutional and the scale issues targeted by this study.

Case Study 2: Inter-disciplinary issues
This case study is concerned with a collaborative group of inter-disciplinary scientists including university and central facility researchers from both chemistry and earth sciences. This study will be based on the use of the ISIS neutron facility (at STFC) [10] and the subsequent modelling of structures based on the raw data, which adds true value to the data. One feature that will be explored is the role of XML for data representation to support easy sharing of the information content of the derived data. Infrastructural components in the process will be identified and workflows modelled emphasising the inter-disciplinary modes.

2.1 Desk Study Synthesis

A desk-based synthesis of existing evidence such as the ABC Advocacy project survey (which itself built on content from earlier StORe, SPECTRa and R4L surveys) and evidence from different structural science domains with a particular focus on chemistry and earth science.

2.2 Data Management Planning Tools

The DCC Data Management Plan checklist [7] will be used to gather information about the planning approaches used by an experimental cohort of chemists and earth scientists in their daily workflows. Particular attention will be paid to *optimising* the application of planning methods to gain *maximum value* from planning effort during all the stages of the Data Lifecycle.

In addition, the Data Audit Framework (DAF) will be implemented within the I2S2 organisational units to establish the status of existing legacy datasets. We will particularly address the potential for re-use of datasets through the application of common standards and data formats, as well as effective data storage, management and curation practice at each site.

2.3 Immersive Studies

The SCARP Project immersive case studies [12] have been highly effective in collecting in-depth "real-world" disciplinary exemplars of data curation practice. We will carry out similar "mini"-immersive studies at Southampton, STFC and Cambridge in order to identify specific requirements.

2.3 Gap Analysis

Gaps in knowledge, practice, infrastructure components and tools will be identified in both the chemistry and earth science domains. Infrastructure components at laboratory, institutional, and large-scale/national levels will be examined to achieve a clearer picture of how vertical integration or scale, can be most effectively achieved.

3. Findings

As mentioned above, several techniques were employed during the requirements gathering phase of the project; the results of which are presented below.

3.1 Desk Study Synthesis

Issues relating to the management of research data have received much attention over the last few years, with a large number of surveys and reports currently circulating in the public domain. The following is a selected list of recent publications considered in the desk study in order to provide background and context to the work being done in I2S2:

- *Sustainable economics for a digital planet: Ensuring long term access to digital information*, Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, February 2010
- *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long term Viability. A comparative review based on sixteen case studies*, A report commissioned by the DCC and SCARP Project, Key Perspectives Ltd, 18th January 2010
- *Communicating Chemistry*, Commentary, Theresa Velden and Carl Lagoze, Nature Chemistry Vol. 1, Dec 2009
- *ParseInsight (Insight into digital preservation of research output in Europe)*, Survey Report, 9th Dec 2009
- *Open Science at Web-Scale: Optimising Participation and Predictive Potential*, Consultative Report to JISC and DCC, Liz Lyon, 6th November 2009
- *Patterns of information use and exchange: case studies of researchers in the life sciences.* A report by the Research Information Network and the British Library, November 2009
- *Chemistry for the Next Decade and Beyond, International Perceptions of the UK Chemistry Research Base*, International Review of UK Chemistry Research, 19 - 24 April 2009, EPSRC
- *Advocacy to benefit from changes: Scholarly communications discipline-based advocacy*, Final report prepared for JISC Scholarly Communications Group by Lara

Burns, Nicki Dennis, Deborah Kahn and Bill Town, Publishing Directions, 9th April 2009
- *Harnessing the Power of Digital Data for Science and Society*, Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council, Jan 2009
- *The UK Research Data Feasibility Study*, Report and Recommendations to HEFCE, UKRDS,19th Dec 2008
- *The Data Imperative*, Managing the UK's research data for future use, UKRDS
- *Infrastructure Planning and Curation, A Comparative Study of International approaches to enabling the sharing of Research Data*, Prepared by Raivo Ruusalepp for the JISC and DCC, 30th November 2008
- *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs*, Report commissioned by the Research Information Network (RIN), June 2008
- *Stewardship of digital research data: a framework of principles and guidelines*, Responsibilities of research institutions and funders, data managers, learned societies and publishers, RIN, Jan 2008
- *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, Consultancy Report to JISC, Liz Lyon, 19th June 2007

A considerably abridged summary of the findings of these reports and surveys indicate that:
- Research teams capture, manage, discuss and disseminate their data in relative isolation with highly fragmented data infrastructures and poorly integrated software applications.
- Conventional systems of publication lead to insufficient information relating to the provenance of results and irreproducible experiments.
- The processes for recognition and reward 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.

3.2 Data Management Planning Tools

The RDMI support project on data management planning tools was set up to help projects within the RDMI strand to use and implement various instruments including:

- Data Audit/Asset Framework (DAF)
- Data Management Plan (DMP) Checklist
- Assessing Institutional Data Assets (AIDA) Toolkit
- Digital Repository Audit Method Based On Risk Assessment (DRAMBORA)
- Life Cycle Information for e-Literature (LIFE)
- Keeping Research Data Safe Surveys (KRDS 1 & 2)

The tools are largely aimed at an institutional context and while suitable for some RDMI projects, they do not fit very well into the context of the I2S2 project which is looking at research and associated processes across institutional boundaries. In addition, many of the tools are "heavy-weight" and would be time and resource intensive to implement fully and formally. We have therefore chosen to informally draw on particular aspects of some of the above tools in the requirements gathering process. In addition, there is specific work relating to cost/benefit analysis in the project (D2.1 Extended cost model; D2.2 Cost analysis phase 1; D4.1 Cost analysis phase 2 and D4.2 Benefits report and business model) which builds on the KRDS 1 & 2 surveys. As a means of raising awareness of data management issues as well as identifying detailed requirements, the DMP checklist template [7] was circulated to research scientists at Cambridge and Southampton.

3.3 Immersive Studies

Immersive studies were conducted between November 2009 and March 2010, centred around the two case studies described in section 2, with a particular focus on the interface between local laboratories and large-scale facilities. The visits were facilitated by Simon Coles (Southampton, NCS); Martin Dove (Cambridge, Earth Sciences) and Peter Murray-Rust (Cambridge, Chemistry). The studies comprised visits to local laboratories followed by visits to the large scale facilities at STFC (DLS and ISIS):

Visit Simon Coles @ NCS 17th Nov 2009
Visit Martin Dove @ Cambridge (Earth Science) 24th Nov 2009
Visit Martin Dove @ ISIS 7th & 14th Dec 2009 (excluding ISIS User Office)
Visit Simon Coles @ DLS 15th Jan 2010 (including DLS User Office)
Visit Peter Murray-Rust @ Cambridge (Chemistry) 4th Mar 2010

Critical to developing an effective data management infrastructure is a thorough understanding of the data as well as the workflows and processes involved in generating and processing them. It is clear that the processes and workflows in each scientific laboratory differ considerably and that a key requirement is an understanding of the file formats in use as well as the inter-relationships between processing software and data files. Consequently, a familiarisation with the details of workflows, processes, software, and file formats was necessary through several one-day immersive studies.

3.3.1 EPSRC NCS & DLS

The EPSRC UK National Crystallography Service (NCS) [3] is an amalgamation of resources at two centres; laboratory-based facilities in the Chemical Crystallography Laboratory at the School of Chemistry, University of Southampton, together with provision of a synchrotron-based facility on station I19 at the Diamond Light Source (DLS) [13]. The NCS operates nationally across institutions and offers two type of experiment service to its users:

Full Structure Determination: where the NCS generates raw data and works up derived data into results. The sample status service will automatically inform users of the successful outcome of a structure determination by means of a computer generated e-mail. The results are sent electronically via e-mail as MS-WORD documents (tables with details of data collection and structure refinement, bond lengths and angles and Figures). Alternatively, these results can be downloaded via the interactive sample status service. The researcher's sample will then be returned in the post.

Data Collection Service: where NCS collects the raw data and turns it into the first stage of derived data. This derived data is then sent to users and they work it up into results. Users of the Data Collection Service can receive their data via e-mail along with a summary sheet (in HTML format) describing details of the data collection. Alternatively the dataset can be downloaded via the interactive sample status service. The sample will be returned to the researcher by post. The data collection summary, along with details about the data collection strategy provides sufficient information to write up a structure for publication.

In addition to routine structure determinations and data collections, the NCS has the capabilities and expertise to: handle extremely small crystals with poor diffraction; deal with twinned and multiple crystals in many cases; handle air and moisture sensitive samples; handle low melting compounds; carry out rapid data collections (particularly useful when the sample deteriorates quickly in the X-ray beam even at low temperatures or for analytical purposes); record multiple data sets in order to follow thermal behaviour, phase transition, reactivity etc.; carry out accurate, high resolution data collections for charge density

determinations; face indexing of crystals; measure diffraction patterns for single crystal/powder composites.

Organisational and Administrative Procedures @ NCS

The NCS provides a service to those eligible to seek research grant support from the EPSRC Chemistry Program. Calls for submission of requests for allocations on the Service are normally issued once a year around August/September to Heads of Chemistry and Subject-related Departments with a closing date usually around early October. Applications are vetted by the Management Advisory Panel and approved allocations announced normally by the end of October.

The standard procedure is to award running allocations valid for a period of one year. For potential users who find an urgent need for access to the Service in mid session, a stream-lined procedure is available whereby a small allocation may be made for the specified project.

Eligible persons who would like to apply for an allocation on the Service may fill in an online allocation web-form. An offline application form is also available and must be accompanied by a 1-page case for support (scientific program) sent either as an e-mail attachment or by post. A confirmation is sent by e-mail on receipt of the application, which will then be sent to the management panel for approval. Once an application has been approved the submitter is sent the appropriate submission forms, including an Experiment Risk Assessment (ERA) form; these forms are returned to the NCS together with a sample of the substance on which the experiment is to be conducted.

When the sample and forms are received, a copy of the ERA is filed in chronological order whilst the original and an additional copy stay with the sample. The submitter of the proposal and sample is issued a digital certificate (generated manually) to allow the progress of the experiment to be tracked.

The NCS operates a sophisticated and dynamic scheduling system whereby priority allocations by the management panel as well as sample rankings from the submitter are taken into account.

During data collection and processing an NCS crystallographer is likely to annotate a copy of the ERA to provide additional information for the submitter. Both the original and annotated sheets together with the sample are returned to the researcher on completion of the experiment. In addition, the submitter is notified of the failure or success of the experiment and provided with the results data in the form of a .zip archive.

Data Generation & Collection @ NCS

The instruments and the current configuration in use are described on the NCS website [3], they include a pair of Bruker-Nonius KappaCCD diffractometers, located on opposite windows of a Bruker-Nonius FR591 rotating anode X-ray generator, see Figure 2.

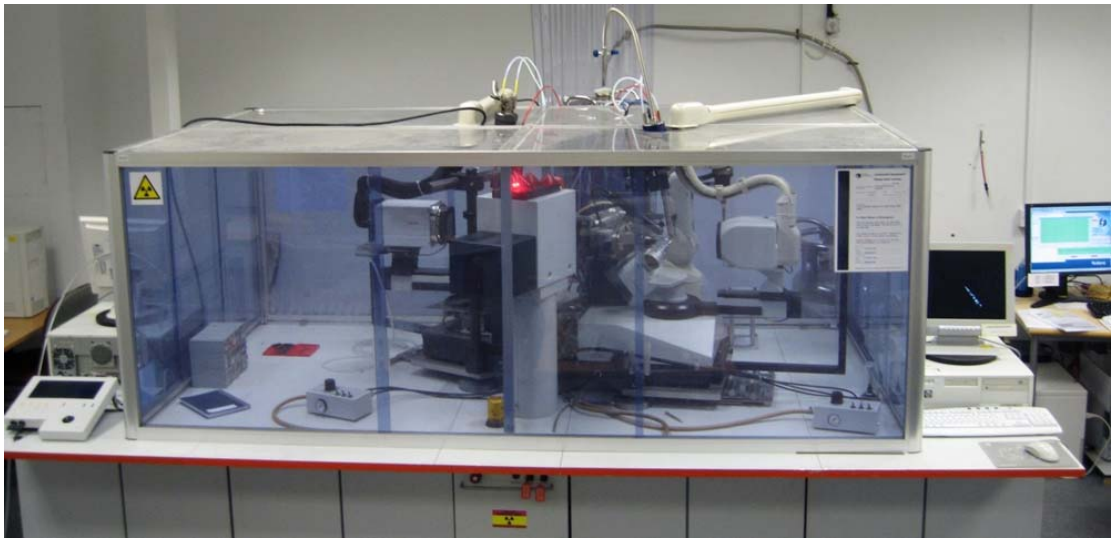

Figure 2: Instrument assembly at NCS [3]

The Service also makes use of a low temperature device, comprising a pair of Cobra© devices and a non-liquid Nitrogen cryostream from Oxford Cryosystems, removing the need for liquid nitrogen in the laboratory. Data may be collected in the temperature range 80 K to 500 K. Software has been developed by the NCS that enables rapid and totally hands-free collection of sets of variable temperature data in this range.

The processes of data collection and reduction are automated using a software application called COLLECT. Following an initial 'pre-scan' to check crystal quality an attempt is made to index the unit cell. Once the unit cell and orientation matrix have been obtained COLLECT will calculate a data collection strategy to access all the reflections in the asymmetric unit. Following collection, the data are integrated using a software tool called DENZO and passed through SCALEPACK. No scaling is actually carried out at this point but the cell is refined using all reflections. An empirical correction for absorption is applied using SADABS, another software tool. An alternative data reduction procedure may be used in the case of twins or other difficult non standard crystals. This revolves around the phi/chi experiment which enables the identification of twin lattices. The program EvalCCD is then used to resolve the overlaps and integrate the reflections writing either an 'HKLF 4' file for the major component or an 'HKLF 5' for both components.

Crystal Structure Determination Workflow @ NCS

Procedures at the NCS indicate that a number of well-defined, sequential stages are readily identifiable and result in a workflow as shown in Figure 3 (note that the pipeline is not entirely linear and does in fact contain several iterative cycles which are not indicated in the diagram). At each stage, an instrument or computational process produces an output, saved as one or more data files which provide input to the next stage. The output files vary in format, they range from images to highly-structured data expressed in textual form; the corresponding file extension names are well-established in the field. Some files also contain metadata, such as validation parameters, about the molecules or experimental procedures.

During the work-up of the data, they progress from being in a state of raw to derived, to final results data. The data collection stage provides JPEG files as representations of the raw data, which are derived from proprietary formats generated natively by the instrumentation used for the experiment. This stage may also have an HTML report file associated with it, providing information relating to machine calibrations and actions and as well as metadata describing how the data were processed. A significant result of the processing stage (process and correction of images) is a standardised ASCII text file {.hkl}, which has become a historical

de facto standard within the crystallography community through its requirement by the SHELXL software (a suite of programs for crystal structure determination from single-crystal diffraction data).

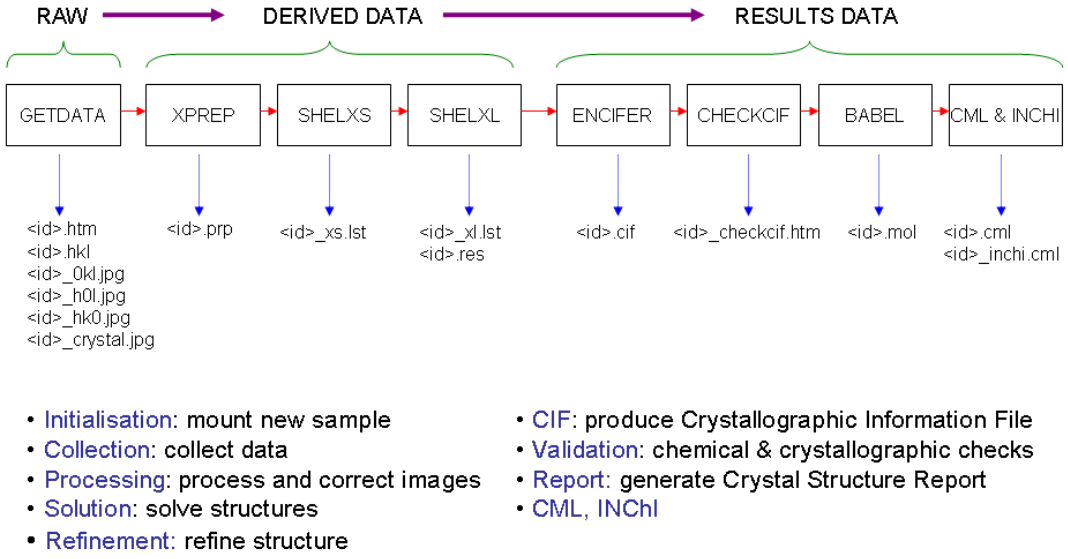

Figure 3: EPSRC NCS Crystallography Workflow

Solving the structures results in a log file {.lst} comprising information relating to the computer processes that have been run on the data by the SHELXS software and a free-format ASCII text file {.prp}, which is generated by software (XPREP). The SHELXL software produces both an output {.res} and a log file {.lst} in ASCII text format as a result of the data work-up process (this is an iterative refinement of many cycles and the output of the final stage is provided in the repository record). There are approximately six versions of SHELXS and SHELXL, which are in use by 80-90% of the community. SHELXS and SHELXL are both commercially and openly available and currently being redeveloped.

The derived data are then converted to results data in the form of the Crystallography Information File (CIF) format [14], which is used as an interchange format and is supported by the International Union of Crystallographers (IUCr). CIF is a publishing format as well as being structured and machine-readable; it is capable of describing the whole experiment and related modeling processes. Associated with the CIF format is the checkCIF software that is widely used to validate CIF files both syntactically and for crystallographic integrity; it is made available as an open web service by the IUCr [15].

Another type of file format included in the final results data is a Chemical Markup Language (CML) encoding [16]. The CML file is translated from the CIF and introduces complimentary semantic information such that between them they provide a complete description of the molecule as well as its chemistry. The {.mol} file is a useful intermediate format for producing the InChI [17], a unique text identifier that describes molecules, and is generated from the {.cif} file (note that the InChi can also be expressed as a URI in the info:inchi

namespace). The file format conversions are performed according to well defined standards using the OpenBabel [18] software obtainable from SourceForge.

Data Storage and Management

Since the NCS is a national service there is felt to be an obligation to retain data, although exactly how long for is not currently specified. The NCS maintains separate strategies for the storage and management of raw data (which is proprietary in nature since it is dependent on specific instrumentation) and, derived and results data which ends up in a normalised, de facto community standard format which is portable and usable by other crystallographers.

Raw data generated in-house at the NCS is stored and preserved in perpetuity off-site at the ATLAS Data Store (Rutherford Appleton Laboratory), an out-sourced service provided by the STFC which currently costs the NCS around £1200 annually. A software script written in-house is used to transfer a .zip archive of the raw and reduced data to ATLAS. Raw data on the ATLAS Data Store goes back to 2002, whilst raw data from 1998-2002 is stored on USB disks stored in the NCS laboratory (migrated from CD's written at the time of generation).

Refined or reduced data is placed in a staging area (currently for one month) from which it is transferred to a laptop for processing by an NCS crystallographer as described above for full structure determination or for a sanity check in the case of data collection only.

Processed data is stored and managed in an institutional data repository (eCrystals [19]), developed to provide open access and rapid dissemination of intermediate, derived and results data from crystallography experiments, as well as linking research data to publications and scholarly communication [20]. eCrystals is constructed on the ePrints.org repository software platform [21] (version eprints-3.0.3-rc-1) which has been customised specifically to cater for crystallography data. A considerable amount of quality and validation checking is performed prior to data files being ingested into eCrystals.

The eCrystals server is managed by a part-time systems administrator with primary training in crystallography. Backups of the repository are kept within the Chemistry department, in another building to where the main server is housed. At the present time the repository comprises 4 terabytes of data; the associated metadata can be exported using a METS profile [22] to allow ingest to an alternative repository platform. The use of OAI-ORE [23] for packaging crystallography data for interoperability purposes is currently under investigation.

Appraisal and Quality Control

Checking, cleaning and refinement of raw data are performed using software. Sanity and validation checks are performed by experienced NCS crystallographers before raw data is transferred to ATLAS. Appraisal, documentation and quality checking are performed before intermediate, derived and results data are uploaded into the eCrystals repository; these include the use of openly available utilities such as *checkCIF* which is maintained by the IUCr [15].

Documentation and Metadata

Raw data transferred to ATLAS does not at present include any metadata or documentation other than filenames.

The eCrystals data repository uses the eBank-UK Metadata Application Profile [24]. This Application Profile (AP) is encoded in the XML schema language (XSD). Broadly speaking, the profile records the following information:

Simple Dublin Core
- *Crystal structure*
- *Title (Systematic IUPAC Name)*
- *Authors*
- *Affiliation*
- *Creation Date*

Qualified Dublin Core (for additional chemical metadata)
Empirical formula
International Chemical Identifier (InChI)
Compound Class and Keywords

The repository uses Digital Object Identifiers [25] as a form of reference identifier as well as the IUPAC International Chemical Identifier (InChi) [17] as a domain identifier.

IPR, Embargo and Access Control

The eCrystals data repository comprises a public and a private part; through the use of an embargo schema, data can be stored in a dark archive and be reviewed periodically for conversion to open access.

Organisational and Administrative Procedures @ DLS (Beam line I19)

A significant proportion of crystalline samples are unsuitable for analysis by standard laboratory equipment, for reasons such as small crystal size or weak scattering. As such, synchrotron radiation provides an invaluable route to the collection of single crystal X-ray diffraction data on such samples, as this provides a source of X-rays at intensities considerably greater than those produced by laboratory sources. In such cases samples are redirected to Beam line I19 at the Diamond Light Source (DLS) [13] and examined by NCS crystallographers during the next available beam time allocation to the NCS.

The DLS User Office is the first port of call for all users of the synchrotron. The necessary procedures and processes are outlined in a Beginner's Guide which is available from the DLS website [10]. In summary, the DLS operates a peer review system for the allocation of beam time for non-proprietary research. The purpose of the peer review is to assess scientific quality in the context of technical feasibility and optimisation of the beam line (each designed to support a particular research community or technique). The review panels rank the proposals according to scientific excellence and technical feasibility and give their recommendations to DLS. The management at DLS makes the final decision on whether a proposal is successful in getting beam time. Decisions regarding approval (or not) are communicated to the applicant by the User Office.

Following approval, the User Office will contact the Principal Investigator (PI) of a successful proposal and direct them to the on-line safety training notes. The PI then needs to inform the Co-Investigators of the necessary arrangements and requirements. The User Office also needs to be notified of the final list of scientists who will be at DLS during the beam time allocation period to enable the preparation of security cards. A welcome pack is made available at the main gate of Rutherford Appleton Laboratory (RAL) on production of photographic identification and includes a security card which enables the user to get access to the site. By default the proximity card will only give users access to the site; they must undertake a safety training test to access the experimental hall, cabins and laboratories. This test must be undertaken every 6 months to ensure users are up-to-date with safety procedures at DLS.

The User Office sends an ERA populated with data from the original proposal to the PI for checking and any amendments. Samples may be brought to DLS by the PI (provided they have been approved by the Safety Office at the proposal stage), but the associated ERAs are submitted (online) by the NCS to the DLS User Office in advance of the scheduled beam time allocation and arrival of NCS crystallographers at beam line I19. The User Office sends the ERAs for health and safety approval, as well as to the local beam line contact.

At the end of beam line allocation time, the PI is required to submit an end-of-beam time survey report. Within three and six months of the experiment the PI is required to submit an

Experiment Report and an Outcomes Report describing results, achievements and any publications resulting from the experiment. The DLS maintains a database of such reports in order to expose links between beam time allocations, research publications and funding sources.

Data Generation & Collection @ DLS (Beam line I19)

Each beam line functions under the management of a Principle Beam line Scientist in association with several dedicated experts who support the PI in setting up the experiment and collecting data from specific beam line instruments.

A detailed description of Beam line I19, is provided on the DLS website; it comprises four sections [13]:

- *Front end* where light is extracted from the storage ring
- *Optics hutch* where certain wavelengths of light are selected and focussed
- *Experimental hutch* housing the experimental equipment; x-rays interact with the sample and are detected using x-ray cameras
- *Control cabin* where the scientific team monitors and controls all aspects of the experiment and collects data.

Although a quick, preliminary check of the raw data is sometimes performed at beam line I19, to assess the quality of the data being collected, the majority of the processing, analysis and solving of structures is normally performed back at the NCS due to time constraints at the beam line itself.

Data Storage and Management

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

The Core Scientific Metadata Model (CSMD) [26] was developed to help organise data derived in investigations using the large-scale facilities at STFC. This model captures the experiment proposal and the acquisition of data, relating data objects to each other and their experimental parameters. The CSMD is currently used in the ICAT suite of tools [27]; a data management infrastructure developed by STFC for the DLS and ISIS facilities. ICAT is primarily intended as a mechanism to store and organise raw data and for facility users to have systematic access to their own data and keep a record for the long term.

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

3.3.2 Earth Sciences, Cambridge & ISIS

The scenario exemplified by Prof. Martin Dove (Cambridge, Earth Sciences) is probably typical of many academic research scientists in the Structural Sciences. In this case, experimentation and data collection are undertaken by a small team of scientists at the ISIS pulsed neutron and muon source (STFC) on the GEM Diffractometer [28].

Organisational and Administrative Procedures @ ISIS

Submission of proposals, peer-review and allocation of beam time, user registration and safety procedures, and end-of-beam time reporting are identical to that at the DLS (see above).

Data Storage and Management

Raw data collected from the GEM instrument is stored at ISIS, once again using the ICAT data management infrastructure as at the DLS.

Processed and derived data from GEM are normally taken off site on laptops or removable drives and sometimes stored on a WebDAV server [29]; the results data are independently worked up by the scientist back at their home institution in Cambridge. The situation is characterised by a lack of shared infrastructure so that data sharing amongst collaborating colleagues is through informal means such as email, ftp and memory sticks; making the management of intermediate, derived and results data a major issue.

Appraisal and quality control, documentation and the addition of metadata as well as IPR and access control are at the discretion of the individual scientist who maintains the data, usually on his or her own laptop, although some programs such as RMCProfile (see below) support and promote the recording of some types of metadata and contextual information.

Crystal Structure Determination Workflow

Figure 4 was developed by Erica Yang based on an original workflow diagram produced by Martin Dove [30]. It shows that as in the NCS case, there are three basic types of scientific data: raw, derived and results datasets. However, there are also a range of other information and control data associated with the production of the datasets. A description of the processes, software applications and tools, as well as human inputs is provided in the case study report which is provided as a supplement to this report [30].

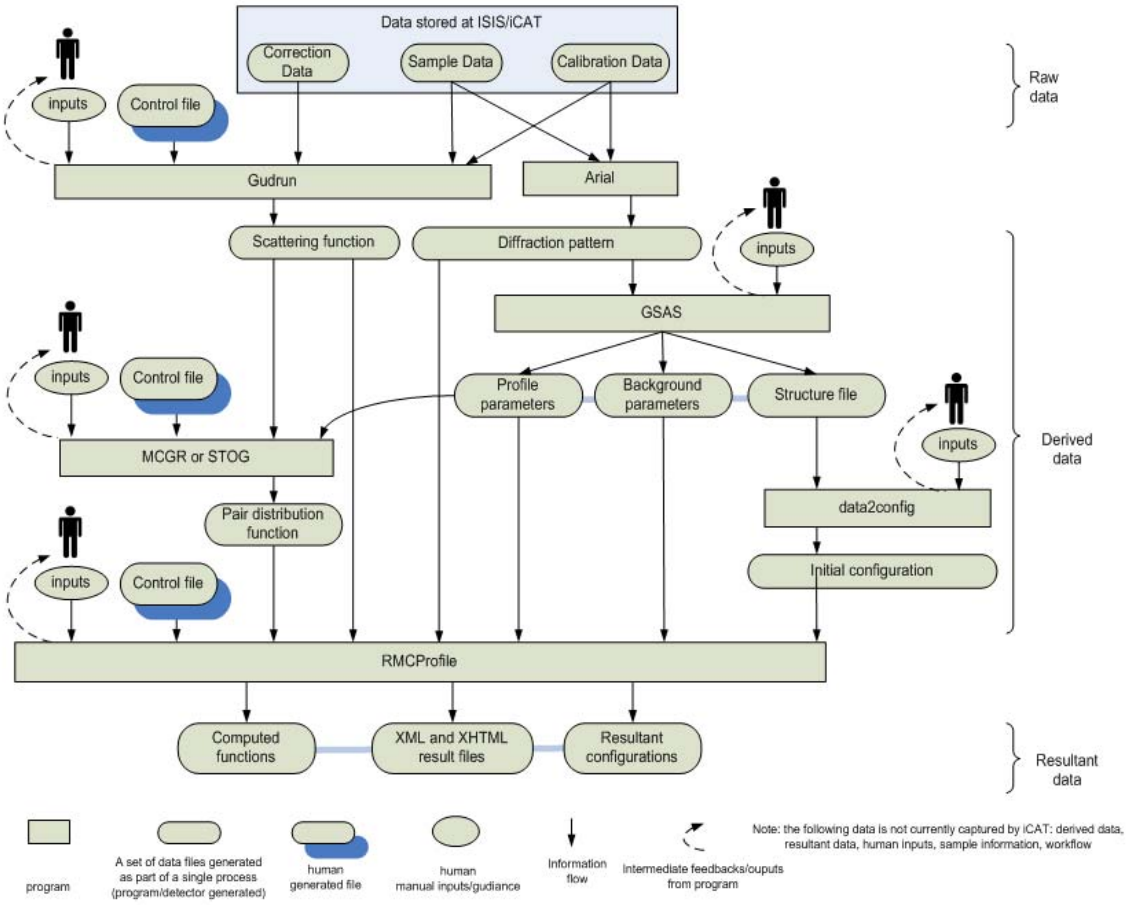

Figure 4: Cambridge Earth Sciences Crystallography Workflow [30]

3.3.3 Chemistry, Cambridge

The Chemistry department at the University of Cambridge [2] consists of a large number of groups covering a spectrum of science, centred on the broad discipline of Chemistry, and ranging from Molecular Biology to Geophysics. The department consists of over 70 academic staff, 110 support staff, 250 postgraduate students and 180 postdoctoral research workers who are supported from central funds or by grants from Research Councils, the European Union, industry, charities or other sources.

The department has a wide range of state-of-the-art instrumentation. Recent acquisitions include 700 and 600 MHz solution state and 400 MHz solid state NMR spectrometers, Q-TOF and FTICR mass spectrometers, single crystal X-ray diffractometers, sub-Angstrom resolution electron microscopes and scanning microscopes.

The specific group we are working with in the I2S2 project is conducting research in the field of Molecular Informatics [31]. Within the project they represent science being undertaken at the research group or departmental level. According to Prof. Peter Murray-Rust, the lead scientist [32]:

“Our research in molecular informatics brings tools from computer science to chemistry, biosciences and earth sciences, integrating humans and machines in managing information.

- We have created Chemical Markup Language (CML), an expanding XML representation of molecular science including molecules, spectra, reactions, computational chemistry and solid state.
- We investigate how computers can be used in communication such as authoring papers/theses. We work closely with several publishers.
- We investigate how the chemical literature can be text and data-mined to discover new science from heterogeneous data sources.
- We are automating the process of computational chemistry by providing expert wrappers to major programs. The results will support *in silico* prediction of molecular and reaction properties for use in safety, pharmaceutical design, enzyme processes.
- We are part of the UK eScience network and are developing the semantic Grid for chemistry. This Grid will seamlessly link databases and services and allow scientists to ask "Google™-like" questions with chemical content.
- We are promoting Open source and Open data and are developing a peer-to-peer system for publishing molecular information at source so it becomes globally available."

Data Storage and Management

The department is currently in the process of enhancing a basic repository which stores crystallographic data through the JISC funded CLARION (Cambridge Laboratory Repository In/Organic Notebooks) Project [33]. The intention is to capture core types of chemistry data and ensure their access and preservation. The department is also implementing a commercial Electronic Laboratory Notebook (ELN) system; CLARION will work closely with the ELN team to create a system for ingesting chemistry data directly into the repository with minimum effort by the researcher.

CLARION will also provide functionality to enable scientists to make selected data available as Open Data for use by people external to the department. The project will use techniques for adding semantic definition to chemical data, including Resource Description Framework (RDF) [34] and Chemical Markup Language (CML) [16]. In addition, the project will address general issues such as ownership of data. Effort will also be put into developing a sustainable business model for operating the repository so that it can be adopted by the department after project completion.

IPR, Embargo and Access Control

One of the major developments in the CLARION Project is that of an embargo manager designed to control the release of data to third party scientists (see Figure 5 below). At present, it is envisaged that this embargo and access system will apply only to results data.

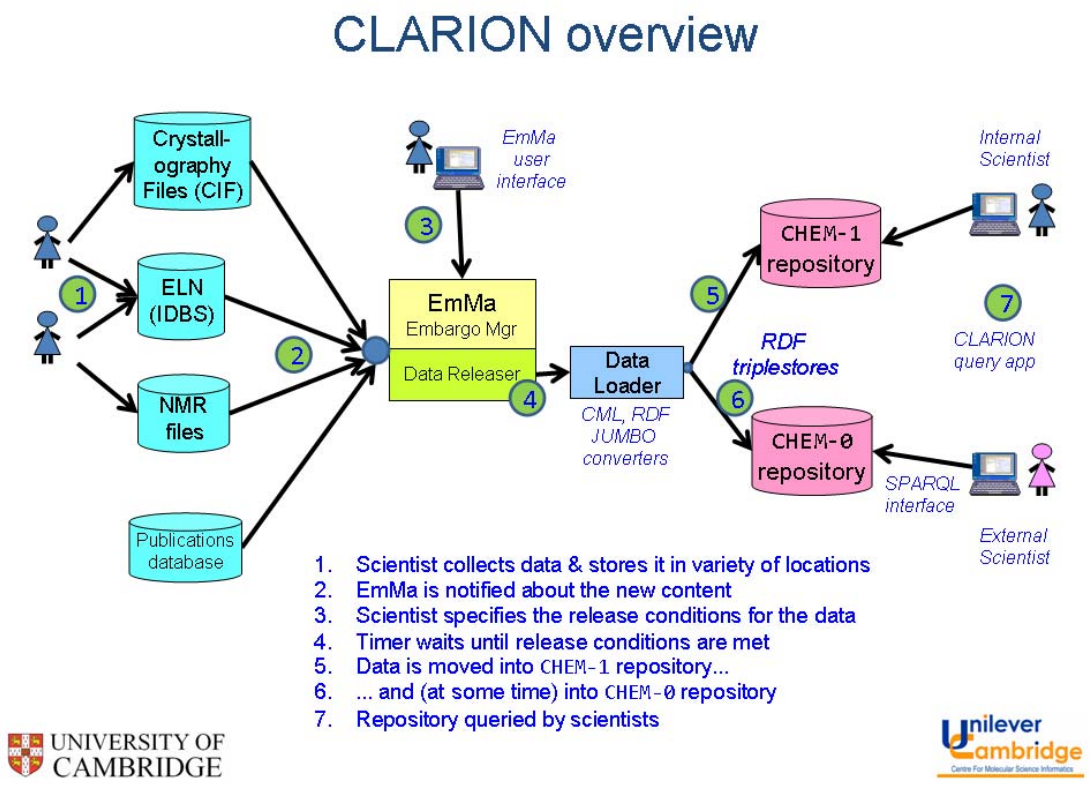

Figure 5: Overview of data management in the CLARION Project [35]

3.3.4 DLS & ISIS, STFC

Large-scale facilities such as DLS and ISIS are operated by STFC on behalf of a huge range and number of scientists from multiple institutions and international communities. As described in sections 3.3.1 and 3.3.2, these facilities have their own processes and administrative procedures including scientific (peer) and technical review of proposals for beam time allocation and the management of user and safety information. They also have dedicated scientists and technical personnel who support the work of visiting scientists.

Data Storage and Management

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

The Core Scientific Metadata Model (CSMD) [26] was developed to help organise data derived in investigations using the large-scale facilities at STFC. This model captures the experiment proposal and the acquisition of data, relating data objects to each other and their experimental parameters. The CSMD is currently used in the ICAT suite of tools [27]; a data management infrastructure developed by STFC for the DLS and ISIS facilities. ICAT is primarily intended as a mechanism to store and organise raw data and for facility users to have systematic access to their own data and keep a record for the longer term.

3.4 Gap Analysis

An idealised and generic scientific research data activity lifecycle model is presented in Figure 6. This model is being developed to serve as a basis for understanding current practice and identifying gaps in local data management processes and procedures at the partner organisations participating in I2S2. It also serves as a point of reference for the cost/benefits analysis work in WP2.

3.4.1 An Idealised Scientific Research Activity Lifecycle Model

The model represents the processes and phases of a typical science experiment project as indicated in the blue text boxes. The stages include: development of the research proposal; its peer-review; carrying out of the experiment; and processing, analysis and interpretation of the data which is eventually reported and published in various forms as research outputs. In addition to these familiar phases, we have incorporated several idealised stages (green boxes) to cater for the long-term management and availability of scientific data; these include appraisal and quality control; documentation including metadata and contextual information; storage, archive, preservation and curation; and IPR, embargo and access control.

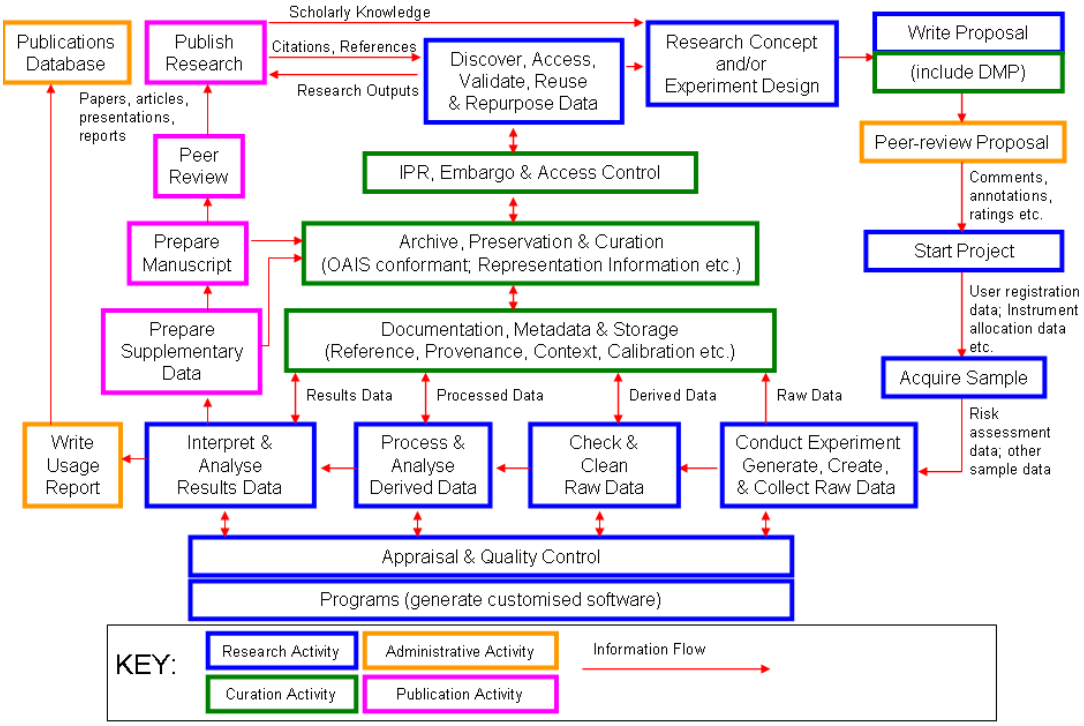

Figure 6: An Idealised Scientific Research Activity Lifecycle Model

It is important to realise that the life expectancy of scientific data has increased over the years as more and more scientific research becomes derivative in nature, dependent on data generated, managed and made widely accessible to third parties. However, effective reuse and repurposing of data requires much more information than the dataset alone. Trust and a thorough understanding of the data is a precursor to its reuse and this in turn necessitates transparency and access to considerable contextual information regarding how the data was generated, processed, analysed and managed.

Consequently, within the I2S2 project, research data is considered to be not only the raw images and numerical datasets that are generated and collected from scientific experiments,

but also the broader categories of information that are associated with such data. These include research and experiment proposals; the results of the peer-review process; laboratory notebooks; equipment configuration and calibration data; processing software and associated control parameters; wikis; blogs; metadata (context, provenance etc.) and other documentation necessary for the interpretation and understanding of the scientific data (semantics); as well as the administrative and safety data that accompany experiments (from the yellow text boxes in Figure 6).

Much of this type of information can be considered to be *Representation Information* (RI), a key concept which underlies the Open Archival Information System (OAIS) Reference Model [36]. RI is a very broad concept, encompassing any information required to render, process, interpret, use and understand data. For example, it may include a technical specification, or a data dictionary exposing semantics or even a software processing tool. An investigation of RI for chemical crystallography data was undertaken in the eCrystals Federation Project [37].

The data and information flows between and into the various phases of the research cycle are represented by the arrows between the boxes. Information generated during the development of the research concept and experiment design is most likely to be textual and paper-based in the form of hand-written notes or hand-drawn diagrams. An electronic record is likely to emerge at the proposal writing stage in the form of a document, or a collaborative tool such as a wiki or blog (or even email). The peer-review process generates additional (most likely textual) information, but also ratings of research proposals and possibly funding and resource allocation data.

Before an experiment can take place, administrative and sample safety information is required to be recorded in the form of either digital or analogue forms which must be checked and approved. An important part of the experiment setup phase is the recording of equipment configuration and calibration information, which is often generated automatically by the instrument being used. Whilst the experiment is in progress, the scientist is likely to record additional information in an analogue laboratory notebook or a laptop computer.

Following the collection of data from the experiment, there is normally a stage in which the raw datasets are cleaned and checked to make them usable, resulting in processed data. Once again, it is likely that the scientist will record any observations or issues in a notebook or laptop.

Results data are produced from an iterative cycle of processing and analysing derived data using software applications and tools. This is the stage in which the real scientific work is performed and it is crucial that adequate records are kept for future reference; most researchers record such information in a laboratory notebook or a laptop.

The traditional end to a scientific research experiment or study is the writing up and publishing of the results in a conference or journal article, with the results data distilled and interpreted to the extent that they cannot be easily checked or verified.

Until recently, it has been the norm that scientists are prepared to share the results dataset with selective colleagues. However, with improvements in technology there is an increasing demand to make available raw, processed and derived data for validation and reanalysis purposes, necessitating data management of these types of data as well as the results data.

Documentation of all types of scientific data (raw, reduced, processed, derived and results) in terms of providing adequate metadata, contextual information and Representation Information, becomes very important for their maintenance and management as well as for additional purposes such as: referencing and citation; authenticity; integrity; discovery and access; search and retrieval; reuse and repurposing; preservation and curation; IPR, embargo

and access management. All of these functions support the scholarly research and communications process (pink boxes in Figure 6 above).

4. Requirements Analysis

It should be borne in mind that this report represents preliminary results and that the specific details of requirements at differing scales of science are likely to become more apparent as the project proceeds in the implementation of the use cases and the associated pilot data management infrastructures.

4.1 Earth Sciences, Cambridge

In the case of Earth Sciences at the University of Cambridge, it is apparent that the greatest need is for 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 (not precluding the sharing of data with external scientists) is also a primary concern as is a requirement for access to research data in the long run so that a researcher (or another team member) can return to and validate the results well into the future.

Consequently, there is a need for basic department or research group level data storage, backup and management facilities which would in addition help to capture, manage and maintain:

- Metadata and contextual information (including provenance)
- Control files and parameters
- Versioning information
- Processing software
- Workflow for a particular analysis
- Derived and results data
- Links between all the datasets relating to a specific experiment or analysis

In addition, any changes should be easily incorporated into the scientist's current workflow and be as un-intrusive as possible.

4.2 Chemistry, Cambridge

In terms of the Chemistry department at the University of Cambridge, the implementation and enhancement of a repository for crystallography data is already underway as part of the CLARION Project. However, this solution will require additional effort to convert it into a robust service level infrastructure.

Over and above the requirements outlined in section 4.1, the situation at the Chemistry department at Cambridge indicates that there is a real need for IPR, embargo and access control to facilitate the controlled release of scientific research data. They have also recognised that valuable information tends to be recorded in laboratory notebooks which are difficult to share and reuse unless stored digitally. The work being undertaken also highlights the importance of data formats and encodings (RDF, CML) to maximise the potential for data reuse and repurposing.

4.3 EPSRC NCS

Since the NCS is a national service there is felt to be an obligation to retain both scientific and administrative data. Data resulting from administrative and safety functions is currently managed and maintained using a hybrid system involving both automated services and a labour-intensive paper-based records-keeping system. Also, paper copies of ERA forms are

annotated by NCS crystallographers and photocopied several times over. In addition, the NCS currently operates a paper-based system for scheduling experiment runs. All of these areas would benefit from further online processing and automation.

The NCS as a case study highlights the issue of referencing data using persistent identifiers. At present, the NCS has to deal with several identifiers per sample (researcher assigned; researcher institution assigned, NCS assigned). On top of these multiple identifiers, an additional identifier (based on the beam line number) is assigned by the DLS if the sample needs to be sent to beam line I19 for processing.

In addition, there is a need to streamline the administrative functions between the NCS and the DLS, for example ERA forms are not currently standardised and therefore require manual intervention.

All data generated through crystallography experiments is considered to have long-term value. It is clear that the eCrystals data repository will require formulation of long-term commitments and objectives with regard to deposit agreements as well as expected services. However, it is recognised that making policy commitments is difficult in an academic environment, which operates under a régime of short-term contracts and funding cycles. In addition, it is worth bearing in mind that formal commitments may well entail legal liabilities. In this respect it is important to secure adequate backing from the host institution, in this case the University of Southampton.

For journal publications that report and link to crystal structure determinations presented in the repository, it is important to satisfy both publishers and the public that the eCrystals repository will have the same stability and longevity as journal publications. At present there is basic information with regard to the contents of the repository, use of the data and citation thereof [19]. A formal preservation policy and strategy is not currently in force and is an area that needs to be addressed.

In the case of crystallography data, it is clear that processing software plays a very important part in crystal structure determination. In particular, software such as the SHELXL/S suite of programs, as well as those for checking and validating CIF files (checkCIF) may also need to be curated and preserved.

Examination of the processes and procedures at the NCS have also revealed the importance of standardised file formats and encoding schemes, since the data must be in a form suitable for the end-scientist to be able to take it away and work with it independently of the services at the NCS.

4.4 STFC

Understandably, at present the considerable infrastructure of the STFC and its large-scale facilities is geared up to storing and managing only the raw data that is generated in-house; a service function implying an obligation to retain raw data in perpetuity. Accepting a commitment to manage derived and results data may lead to additional issues with regard to IPR and ownership since these types of data result from the application of a scientist's knowledge and expertise.

However, as the case studies above demonstrate there are efficiencies and benefits to be gained by working across organisational boundaries through an integrated approach. For example, the use of standardised ERA forms and unique persistent identifiers would considerably simplify inter-organisational communications and the tracking, referencing and citation of datasets.

The CSMD and its implementation in ICAT currently serve the purposes of the STFC very well, making it a good candidate for further development and extension to take account of the

wider scientific experiment scene and incorporating the needs of organisations outside of the STFC. The model would need to be extended to cater for several areas such as curation and preservation.

5. Requirements Synthesis

The requirements gathering process has revealed that workflows, scientific and administrative processes and practices vary considerably between the research laboratories and organisations examined within the I2S2 project. Furthermore, at present, the aforementioned workflows, processes and practices are very poorly captured and documented (if at all). It should be noted that there are also variations in the terminology used between differing laboratories and services as well as between disciplines. For example, the terms used to describe datasets at differing stages of the data capture and processing pipeline (e.g. raw, reduced, processed, derived and results) evoke differing meanings in different laboratories and disciplines.

The four broadly defined levels of research science examined (individual researcher, team, and medium-level service to large-scale facility) reveal the huge diversity of requirements depending on the situation, circumstances and level of data management infrastructure currently in place. Furthermore, there will undoubtedly be additional discipline specific differences and practices which will emerge as the project progresses to pilot implementation stage.

It is apparent that the requirements range from basic storage and backup facilities (at least initially) to much more sophisticated needs such as embargo control and the structuring and linking together of data. These requirements need to be viewed in the context of a research group, a department, an institution or a disciplinary community.

One way of reconciling this huge diversity in requirements is to try to identify similarities and differences and thereby formulate an integrated approach to the development of a data management infrastructure. Recording adequate metadata and contextual information is critical to supporting an efficient scholarly communications process based on data-driven science and in particular to facilitate the following:

- Maintenance and management of data
- Linking together of all data associated with an experiment
- Referencing and citation of datasets
- Authenticity validation
- Integrity control
- Provenance
- Discovery, access, search and retrieval
- Preservation and curation
- IPR, embargo and access management
- Interoperability and data exchange

In addition, how the data itself is structured determines how reusable it is. Open and linked data promotes research across disciplinary boundaries [38].

The use of standards is another important area serving the need for interoperability, data reuse and repurposing. To maximize the potential of research data, it is necessary to have and encourage the use of standards for:

- Search, retrieval and access
- File formats
- Data capture, processing and publishing
- Recording metadata and contextual information

For implementing solutions to the requirements identified in an integrated manner so as to span organizational boundaries, some relevant technologies include:

- Persistent Identifiers (URIs, DOIs etc.)
- Metadata schema (PREMIS, XML, CML, RDF)
- Controlled vocabularies (ontologies)
- Integrated information model (structured, linked data)
- Extensions to CSMD & ICAT
- Interoperability and exchange (OAI-PMH, file formats)
- Data packaging (OAI-ORE)
- OAIS Representation Information

In the introduction to this report, we included human elements as being a part of the data management infrastructure; cultural issues play a significant part in the adoption and success or otherwise of a technical solution and should not be under-estimated. Although not within the scope of the I2S2 Project, we nevertheless recognise the importance of advocacy, best practice guidelines and training for researchers. Awareness should be raised of the responsibilities that manifest themselves at different roles and levels of scale (research student, research supervisor, research laboratory, department and institution as well as service facilities at regional, national and international levels).

6. Conclusions

Despite the considerable variation and diversity in requirements between the different scales of science being undertaken (individual research scientists and service facilities), a relatively common thread has become apparent in the form of a need to be able to manage all data (as defined in the broadest sense in section 3.4.1) as they are collected, generated and processed during the course of research experiments.

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); making it onerous for researchers to mange 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.

We conclude that there is merit in adopting an integrated approach which caters for all scales of science (although the granularity and level of integration is an area that needs further investigation). Furthermore, an integrated approach to providing data management infrastructure would enable an efficient exchange and reuse of data across disciplinary boundaries; the aggregation and/or cross-searching of related datasets; and data mining to identify patterns or trends in research and experiment results.

In addition, demands are now surfacing for “Open Methodology”, such that making data alone openly available is insufficient and there are now expectations that the methodologies used in processing and analysing them should also be made readily accessible. The work being undertaken in the I2S2 project (in terms of building a robust data management infrastructure) has the potential to form a foundation on which differing methodologies can be both run and exposed to third parties for easier sharing.

References

(URLs checked and accessed 7th July 2010)

1. Earth Sciences, University of Cambridge, http://www.esc.cam.ac.uk/
2. Chemistry, University of Cambridge, http://www.ch.cam.ac.uk/
3. EPSRC National Crystallography Service (NCS), University of Southampton, http://www.ncs.chem.soton.ac.uk/
4. Science & Technology Facilities Council (STFC), http://www.stfc.ac.uk/home.aspx
5. I2S2 Project Plan, http://www.ukoln.ac.uk/projects/I2S2/
6. Data Audit Framework, Digital Curation Centre, http://www.data-audit.eu/
7. Data Management Plan Checklist (DMP), Digital Curation Centre, http://www.dcc.ac.uk/news/dcc-data-management-plan-content-checklist-draft-template-public-consultation
8. Keeping Research Data Safe ((KRDS 1), JISC, http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx
9. Keeping Research Data Safe Phase 2, JISC, http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx
10. The DIAMOND Light Source (DLS), STFC, http://www.diamond.ac.uk/
11. ISIS, STFC, http://www.isis.stfc.ac.uk/index.html
12. DCC SCARP Project, http://www.dcc.ac.uk/projects/scarp
13. Beamline I19, Small molecule single crystal diffraction, http://www.diamond.ac.uk/Home/Beamlines/I19/layout.html
14. Crystallography Information File (CIF) format, International Union of Crystallographers (IUCr), http://www.iucr.org/resources/cif
15. International Union of Crystallographers (IUCr), http://www.iucr.org/resources/cif
16. Murray-Rust P., Chemical Markup Language - A Simple introduction to Structured Documents, O'Reilly XML.com, http://www.xml.com/pub/a/w3j/s3.rustintro.html
17. International Chemical Identifier (InChi), International Union of Pure and Applied Chemistry (IUPAC), http://www.iupac.org/inchi/
18. Open Babel, The Open Source Chemistry Toolbox, http://openbabel.org/wiki/Main_Page
19. eCrystals archive for Crystal Structures generated by the Southampton Chemical Crystallography Group and the EPSRC UK National Crystallography Service (NCS), NCS, University of Southampton, http://ecrystals.chem.soton.ac.uk/
20. Duke M., Day M., Heery R., Carr L., Coles S.: Enhancing access to research data: the challenge of crystallography, Proceedings JCDL'05, Denver, Colorado, USA, 2005
21. ePrints.org repository software, http://www.eprints.org/
22. Metadata Encoding and Transmission Standard, http://www.loc.gov/standards/mets/
23. Open Archives Initiative, Object Reuse and Exchange (OAi-ORE), http://www.openarchives.org/ore/
24. eBank-UK Metadata Application Profile, http://www.ukoln.ac.uk/projects/ebank-uk/schemas/profile/
25. Digital Object Identifier System, http://www.doi.org/
26. Matthews B., Sufi S., Flannery D., Lerusse L., Griffin T., Gleaves M., Kleese K., Using a Core Scientific Metadata Model in Large-Scale Facilities, 5th International Digital Curation Conference (IDCC 2009), London, UK, 02-04 Dec 2009, http://epubs.cclrc.ac.uk/work-details?w=51838
27. ICAT Project, http://code.google.com/p/icatproject/wiki/IcatMain
28. GEM Diffractometer, ISIS, http://wwwisis2.isis.rl.ac.uk/disordered/gem/gem_home.htm
29. WEBDAV resources, http://webdav.org/
30. Yang E., Martin Dove's RMC Profile Diagram, Internal Report, WP1, I2S2 Project, July 2010
31. Unilever Cambridge, Centre for Molecular Science Informatics, http://www-ucc.ch.cam.ac.uk/
32. Peter Murray-Rust, Staff web page, Chemistry, University of Cambridge, http://www.ch.cam.ac.uk/staff/pm.html
33. Chemical Laboratory Repository In/Organic Notebooks (CLARION) Project, JISC, http://www.jisc.ac.uk/whatwedo/programmes/inf11/sue2/clarion.aspx
34. Resource Description Framework (RDF), W3C, http://www.w3.org/RDF/
35. Brookes B., The CLARION Project, presentation at I2S2 Modelling Workshop, Feb. 2010, STFC Rutherford Appleton Laboratory, UK, http://www.ukoln.ac.uk/projects/I2S2/events/modelling-workshop-2010-feb/

36. 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
37. Patel M., Representation Information for Crystallography Data, WP4, eCrystals Federation Project, 19th May 2009, http://wiki.ecrystals.chem.soton.ac.uk/images/e/e1/ECrystals-WP4-RI-090519.pdf
38. Panton Principles, Principles for open data in science, http://pantonprinciples.org/