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 Implementation Plan	
Reporting Period	*N/A*	
Author(s) & project role	Erica Yang (Senior Research Engineer, RAL, STFC)	
Date	18th October 2010 **Filename** I2S2-WP3-D3.2-ImplementationPlan.doc	
URL		
Access	Public	x General dissemination

D3.2 Implementation Plan

Work Package 3

April 2010 – December 2010

JISC I2S2 Project

Document Details

Author:	Erica Yang (STFC Rutherford Appleton Laboratory)
Date:	18th October 2010
Version:	0.8
File Name:	I2S2-WP3-D3.2-ImplementationPlan.doc
Notes:	

Acknowledgements

Executive Summary

This deliverable describes the implementation plan for the I2S2 project. It comes at an interesting time: it is written *after* we have gained some initial experience of designing and developing a preliminary pilot implementation for capturing, storing, and visualising the derived data generated throughout the analysis pipeline of an exemplar structural science experiment. Such experience is invaluable at this stage of the project as it not only helps us to narrow down our efforts to a few key areas that need most attentions but also give us useful indications on which requirements can be realistically addressed within the timeline of the project.

Specifically, we address *six out of sixteen* findings resulted from requirements gathering process, as reported in the Executive Summary of the Requirement Report, namely:

1. A robust data management infrastructure which supports each researcher in capturing, storing, managing and working with all the data generated during an experiment.
2. Internal sharing of research data amongst collaborating scientists, such as between a PhD student and supervisor.
3. Capture, management and maintenance of:
   (1) Metadata and contextual information (including provenance)
   (2) Control files and parameters
   (3) Versioning information
   (4) Processing software
   (5) Workflow for a particular analysis
   (6) Derived and results data
   (7) Links between all the datasets relating to a specific experiment or analysis
4. Changes should be easily incorporated into the scientist's current workflow and be as un-intrusive as possible.
5. 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.
6. 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.

This deliverable gives the rationales behind why these six areas are chosen (and why the others are not), explain our implementation strategy in each of the chosen areas, and lay down the plans for the implementation for the rest of the project. In retrospect, concentrating our limited resources in the relatively short timeframe of this project has allowed us to make solid steps so that towards the first anniversary of this project, we have a tool that is concrete enough to showcase to our targeted audience – the experimental scientists, to engage them, and to gather the most important feedbacks and valuable suggestions from these people who we are developing the tool for.

Document Revision History		
Version	**Date**	**Comments**
0.1	26th Sept. 2010	Scope and initial outline
0.2	27th Sept. 2010	Set the theme and outline for STFC internal comments
0.3	28th Sept. 2010	Stake holder analysis
0.4	29th Sept. 2010	Implementation plan
0.5	1st Oct. 2010	Add summary
0.6	1st Oct. 2010	Out again for STFC internal comments
0.7	6th Oct. 2010	Incorporated feedbacks from Manjula and revision based on a telco message from Simon
0.7.1	12th Oct. 2010	Draft out to project mailing list asking for feedbacks
0.7.2	15th Oct. 2010	Draft out to STFC internal for comments
0.8	18th Oct. 2010	Draft out to project mailing list asking for comments and feedbacks
1.0	6th April 2011	Final version

Contents

1. Introduction

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

"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."

These outputs are drawn from sixteen major requirements raise by four groups of stakeholders, namely, individual researchers, university research groups, national service provider, and large facility operator. It is a comprehensive set of requirements, embracing a wide range of activities, from research management, project planning and execution, publishing research outputs, and finally to long term research data preservation and curation. The main aim of the current deliverable is to identify and decide which ones are the most urgent unmet demands and set out the plan to derive and implement the solutions to address these problems.

The rest of this report is organised as follows. Section 2 describes our methodology to narrow down the requirements. As a result of the selection exercise, a set of highly focussed requirements will be examined closely in Section 3, which details our implementation strategy and the timeline for completing the tasks. Section 4 concludes the report.

2. Targeted Requirements

I2S2 is funded under the JISC managing research data programme which has a strong theme of identifying and addressing the unmet demands in the research data arena. In order to address open research data problems, it is important to understand the followings:

- Who are the stakeholders;
- What are their concerns; and
- Why existing solutions do not work.

We believe that understanding the stakeholders is the key. The first part of this section performs a stakeholder analysis, briefly analysing the stakeholders involved in the research lifecycle and the second part of this section delves into the research activities of the stakeholders, aiming to identify the unmet demands in the current research practice.

---

1 Herein, without explicitly stated otherwise, "the report" refers to the deliverable D1.1 Requirements Report [1].

2.1. Stakeholder Analysis

Many different types of stakeholders are involved in the idealised scientific research activity lifecycle model presented in Section 3.4.1 in the requirement report. The model is reproduced here for easy referencing. As colorised in the Figure 1, the activities can be broadly and coarsely classified into four groups: research (blue boxes), administration (orange boxes), archive/curation (green boxes), and publication (purple boxes).
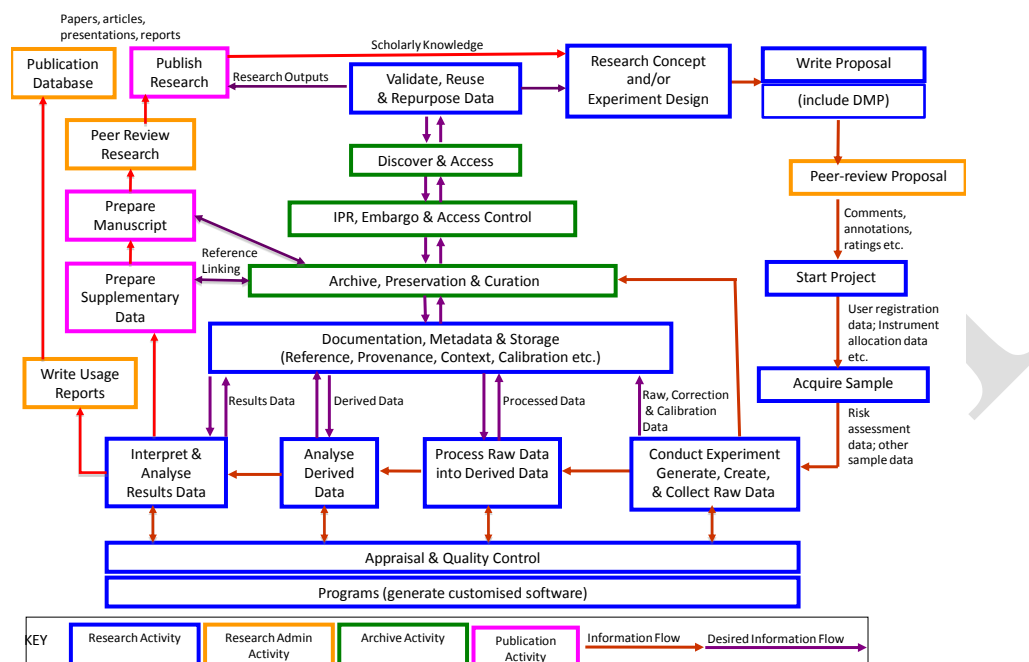

Figure 1 The Idealised Scientific Research Activity Lifecycle Model

Table 1 identifies the stakeholders against each group of the activities. Even at this level of coarse classification, it is evident that researchers have little involvements in the administration and archival/curation activities.

From a researcher's point of view, gaining recognition and accreditation are the perhaps two most important drives. Research activities are the means to obtain support for their research whilst publications are the main output from research which feed back to their research activities. This explains why, in reality, they are not very enthusiastic with the administrative and curation activities as these are perceived as "additional work to their research". In other words, things that are important to administrators or curators may not be as important to researchers.

Table 1 Research Activities vs. Stakeholders

Activity	Stakeholder
Research	University researcher, scientist (may work for service providers or facility operators)
Administration	University department, national service provider, large facility operator
Curation	University department, national service provider, large facility operator

Publication	University researcher, scientist, decision makers

Research activities are the foundation of the entire model because all the other activities are built upon them. This has led to our decision that the unmet demands from researchers should be tackled with the highest priority. Fundamentally, research data are produced by researchers. If researchers cannot do their jobs properly, all the other activities (administration, publication, and curation/archival) down the line will be affected.

The research activities in the lifecycle model can be further classified into three categories: planning, execution, and, result dissemination and repurposing. With this categorisation in mind, we shall now explore closer into each and understand the stakeholders behind each group of activities and their unmet demands.

2.2. Categorising Research Activities

A research team is often comprised of two types of researchers:
- Senior members (professors, research fellows, collaborators)
- Junior members (research assistants, Ph.D. students, master students)

Senior members are often involved in planning, supervising, and publicising research whilst junior members execute the detailed work, although how each type of researchers is involved in the activities and the level of involvements varies significantly from one group to another.

2.2.1. Research Planning

Research Planning embraces the following activities:
- (conceiving) research ideas
- Designing experiments
- Writing proposal
- Starting project

Research planning activities are often carried out by senior research members in a team. How much time a researcher spends on these activities also varies significantly depending on the experience and the research field of the researcher. During one of our project wide workshop, a professor pointed out, "it only took me two hours to write up a proposal. These are pretty standard!" That seems to suggest that these are *not the "key"* activities to senior researcher. Junior researchers often have limited involvements in these activities. Hence, it seems to suggest that there are not much unmet demands in handling research planning activities.

2.2.2. Research Execution

Once funding is secure and people are in place, the "real" research activities begin. For structural sciences, this involves:
- Acquiring and preparing samples
- Conducting experiment
- Gathering raw data
- Analysing and interpreting data

Researchers at all levels are involved in these activities. However, from the in-depth study of the detailed requirements and processes of the Earth Science use case [2], it reveals that researchers spend a significant amount of their time and efforts in dealing with these activities. Based on [2], Table 2 gives a summary of the estimated time spent on each type of activities. The timeframe of a project starts when the funding is started and ends when the funding is finished.

Table 2 Activities vs. Time Spent and Stakeholders

Activity	Time Spent (during a project)
Sample related	Unknown (prior to an experiment)
Conducting experiment (ISIS experiment)	3-4 full days
Gathering raw data	9-16 times throughout an experiment[2]
Handling Data (including cleansing, reducing, processing, and analysing)	Many months

Data Handling and Analysis Pipeline

Data handling is the most time consuming as well as labour intensive process among the activities. It is also a complicated and sometimes daunting process. According to Professor Martin Dove, who has many years of experience in conducting neutron experiments and analysing neutron data, the difficulty of dealing with the raw data is as follows:

"The problem with raw data is even more acute for new and young researchers. You really have to know how to convert the raw data into something useful, and unless you are sufficient of an expert the raw data are completely meaningless. I think that the raw data will simply *frighten* the young researcher."

The analysis is conducted in a highly collaborative fashion, involving a small team of researchers, such as RAs, Ph.D. students, a professor, and an ISIS instrument scientist. People involved are often responsible for a small part of the experiment (often this small part is a program in the analysis pipeline.)

Reproduced from [2], Figure 2 shows a typical data analysis pipeline for the neutron data gathered from one of the ISIS instruments, called GEM, using the Reversed Monte Carlo (RMC) technique. The pipeline appears to be fairly static. But, in a fast changing scientific research field, this is often not really the case. Many programs in the pipeline can be replaced by other functionally equivalent programs, which could be written in a different programming language, with a more efficient algorithm, or simply because it is in a different discipline where researchers use a different set of programs to perform the same job.

Science evolves by learning from the past. This is also true in dealing with scientific data. Scientists learn from their data: scientific programs evolve as scientists progressively develop in-depth understanding of their own research data and the past analysis methodologies. Therefore, any tools developed for capturing research data as well as the associated metadata have to be flexible and adaptable to scientists' workflow. Putting any constrains in their

---

2 Raw data is collected at the end of each run of an experiment. Experiments are conducted in runs, each run lasts between 6 to 8 hours throughout a 24-hour cycle.

workflow or practice, in other words, confining them to a certain types of workflow, will simply reduce the chance of those tools being used productively. What are really needed are tools that can capture the data as research develops.
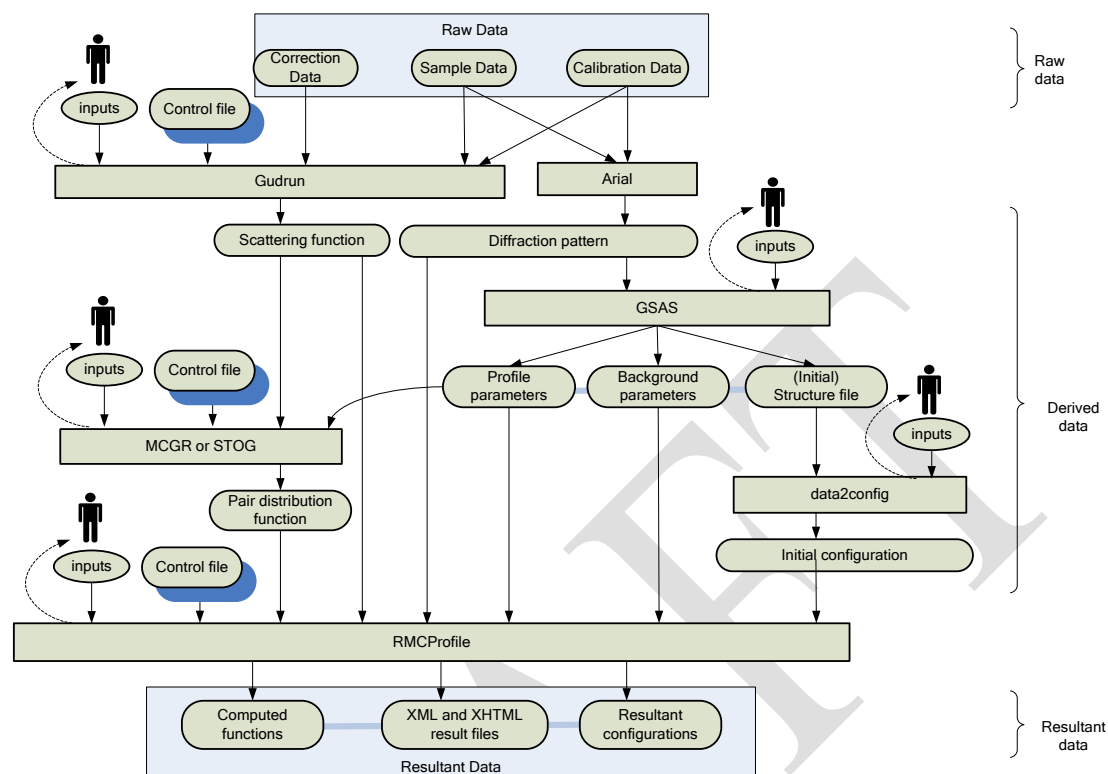

Figure 2 Neutron Data Analysis Using the RMC Method

Data Storage

Processed data is passed on using conventional methods such as, emails, web store, and shared file store. The common practice is to pass on a (physical) *copy* of the derived data for the people responsible for the next program/process in the pipeline to deal with. Good tracking and bookkeeping practice are necessary here to ensure the data flows correctly and promptly from one person to another. However, as all these are currently done by the conventional methods, it is error prone and not effective (cannot scale up to a large collaboration or reuse by other people). Hence, this is clearly an area that a big difference can be made by introducing a shared data management infrastructure.

Additionally, as highlighted in the supplementary report [2], the tools and programs used by the structural sciences community (or more specifically, neutron and synchrotron research as studied in the report) are highly specialised such that they are often written by scientists themselves. For computational scientists, Fortran, C, and C++ are still the favourites. To make things more complicated, these programs can run on a wide range of platforms, VMS, Linux, Unix, and Windows, although some of them may fade out in the coming years. Hence, it is unrealistic to expect the adoption of a common interface among all these well established programs to achieve inter-operability among the programs. Hence, in the foreseeable future, it is still challenging to expect fully automated execution of scientific workflows. From a data management point of view, this means that scientists may still need to manually handle the inputs and outputs of analysis programs.

2.2.3. Research Dissemination

At this stage, the main activities are to produce publications and curate the research outputs (e.g. paper, design, and code) for long term usage. Under the current "publish or perish" research culture, researchers are naturally motivated to produce publications. However, curation is a different matter because it often rises to the agenda towards the end of a research cycle after much of the research work is finished. This often creates problems as the quality of any curation effort depends on the quality and completeness of the input source to the curation process. If the essential metadata, such as provenance or context, are missing from the source, it is difficult to ensure the reusability of the preserved research outputs.

2.3 Identifying the Targets

The main message from the above analysis is that there are clearly unmet demands in managing the analysed (or derived) data for researchers. Hence, we have distilled six out of the sixteen requirements from the report, all of which satisfy two criteria: a) directly related to researchers' data analysis work; b) addressing the requirement can lead to increased of productivity in research work. They are:

1. A robust *data management infrastructure* which supports each researcher in capturing, storing, managing and working with all the data generated during an experiment.
2. *Internal sharing of research data* amongst collaborating scientists, such as between a PhD student and supervisor.
3. *Metadata capture, management and maintenance* of:
   (1) Metadata and contextual information (including provenance)
   (2) Control files and parameters
   (3) Workflow for a particular analysis
   (4) Derived and results data
   (5) Links between all the datasets relating to a specific experiment or analysis
   (6) Capture, management: Processing software
4. Changes should be easily incorporated into the scientist's current workflow and be as *un-intrusive* as possible.
5. 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.

Due to the resource constrain within the timeframe of this project, the following requirement is excluded from the list of requirements that we will address.

1. Department or research group level data storage, backup and management facilities.
2. Sharing of data with third parties.
3. 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.
4. Where crystallography data repositories already exist, there is a requirement to develop them into a robust service incorporating curation and preservation functions.
5. There is a real need for IPR, embargo and access control to facilitate the controlled release of scientific research data.
6. Valuable information commonly stored in analogue laboratory notebooks is difficult to share and reuse and needs to be stored digitally.

7. 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.
8. Paper and or hybrid record-keeping and resource scheduling systems would benefit from automated processing.
9. Use of consistent and persistent identifiers would greatly aid the seamless flow of information between organisations, applications and systems.
10. There is a need to streamline administrative functions between organisations, for example through the use of standardised Experiment Risk Assessment forms (ERAs).
11. 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.
    a. Versioning information
    b. Maintenance of: Processing software

3. Tasks and Timeline

Depicted in Figure 3, the implementation plan consists of three categories of tasks as follows:

Category A. Understanding Requirements runs from the beginning of the project to May 2010. The major outputs from these tasks are the *requirements reports*, which were delivered to JISC on July 2010.

Category B. Pilot Implementation and Development is about the design and development of the *software tools* to address the problems identified in Category A and gather feedbacks from the targeted communities who trial the tools to allow continuously improvements on the tools. Its timeframe runs right after the requirement gathering to the end of the project. As shown in the Figure, this is an iterative process. We feel that, given the relatively short timeframe of the project, this is an effective way to maximize its potential impact by engaging users closely along the software development process. As this is still ongoing, the information presented in this report reflects the current state of the work in this category. The tasks in this category run from the beginning of May till December 2010. Two major deliverables are expected from the tasks in this category: a) the deliverable D3.3 describing the pilot implementations with the details of the design and source code; and b) the sourceforge release of the source code of the tools developed for the pilots. D3.3. will be delivered to JISC at the beginning of 2011 after we wrap up the tasks in December 2010. At the time of writing up this deliverable, we have already put up a preliminary source code release of the tools on the sourceforge repository. This is accompanied by a wiki page describing the tools. We expect to make a formal sourceforge code release with detailed installation, configuration, as well as a demonstration website available in early 2011.

Category C. Information Modelling focuses on the development of the I2S2 information model, which runs, in parallel, to the categories A and B tasks. It starts with a gap analysis of existing metadata standards for scientific data management, progresses into a detailed extension work to capture and accommodate analysed data for scientific data analysis pipeline, and ends with a final report on the information model. We started the tasks in this category in February and expect them to be completed by December 2010. The major deliverable out of the tasks here is the deliverable D3.1 describing the integrated information model, which will be available at the end of December 2010.

Each task in the figure roughly corresponds to a point on the timeline where the task commences. All tasks finish within the time span of the corresponding category. *It should be pointed out that this deliverable focuses on the implementation of the tasks spanning between months 4 – 15, i.e. the tasks in WP3 – Harmonisation and Implementation.* As the major

deliverable from the tasks in Category A has been completed at the time of writing this report, we will focus on Categories B and C in the rest of this section.
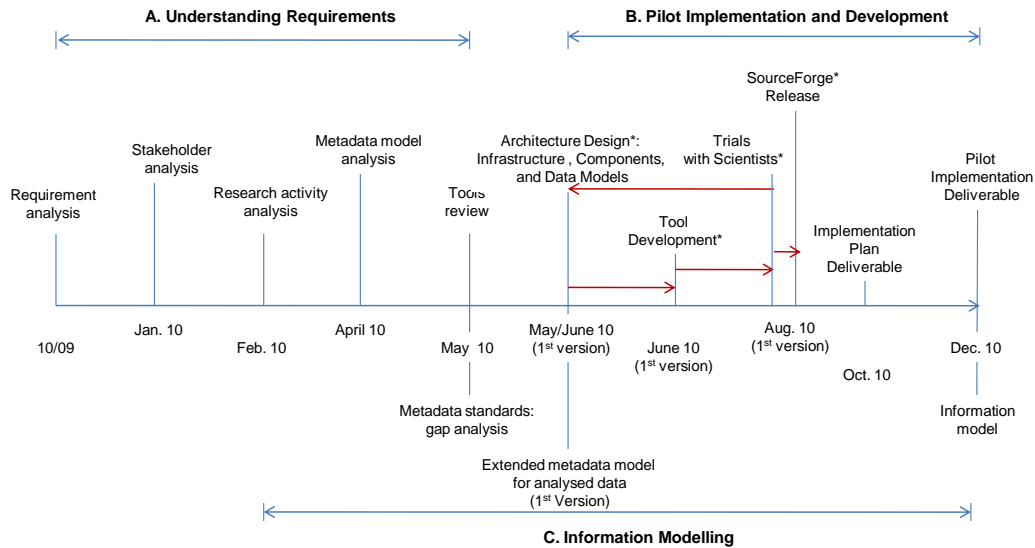
*: these tasks are performed in iterations till the end of the project

Figure 3. Implementation Plan: Tasks and Timescale

4. Pilot Implementation and Development

After justifying the key requirements we are targeting in this project, we shall now describe the category B tasks of pilot implementation and its current progress. This category involves two types of tasks: design and development of tools for data analysis, and documentation of the design and development.

The first version of the pilot implementation, named **ICAT-personal** (earlier, it was referred as **icatlite**) was developed at the end of May and early June 2010. The design and implementation [6] was first presented at the I2S2 internal progress meeting at Bath University. A roadmap for the pilot implementations is depicted in Figure 4, which essentially describes three types of development work:

- BLUE boxes represent the ground work that has been done in the past and will be built upon in this project. The components behind were available prior to the project started. Section 4.1 describes the work in the blue boxes.
- GREEN boxes represent the components that were planned to be built in this project until the end of the WP3, i.e. December 2010. The very first version of some of them also included in the diagram. Section 4.2 describes the infrastructure components that were planned in June 2010. At the time of writing this report, we are well into the development and refinement phases of these components.
- ORANGE boxes represent the I2S2 related components that were being developed (or going to be developed, but *outside the scope* of I2S2) in parallel

to the I2S2 project. For example, there is plan to integrate ICAT-personal with the ICAT infrastructure in the nearer future. This is a related effort, represented as an orange box titled "international facility data repository" in the diagram. Similarly, the orange box, titled "Publication", represents the plan to integrate ICAT-personal with the ePub system, the STFC electronic publication system in the future. It is also a possibility to extend existing STFC eScience internal scientific applications to utilize the data facilities (e.g. search, ingest, and get) made available through ICAT-personal. This potential future effort is represented as an orange box titled "Application embedded client" in the diagram.
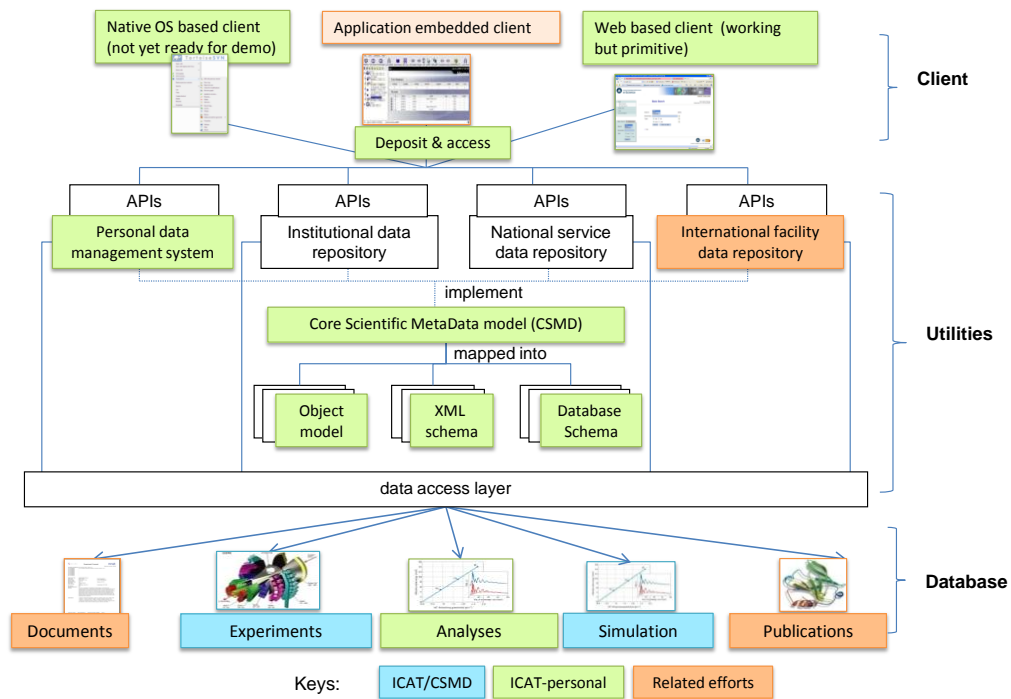

Figure 4. A Roadmap for the Pilot Implementations (June 2010)

4.1. The Baseline of the Pilots

The roadmap illustrates the context of the implementation, which is based on an existing production data management infrastructure of STFC for ISIS and Diamond, called **ICAT** (a short name for **I**nformation **CAT**elog). The metadata model underpinning the ICAT infrastructure is based on an extended version of the **C**ore **S**cientific **M**eta**D**ata model (CSMD) [5]. As illustrated in Figure 4, the current ICAT supports experimental (raw) data, capturing all aspects of a scientific experiment, from sample information, experimenters, purpose of an experiment, experimental conditions (e.g. temperature), and timestamp for each raw data files generated during an experiment. These "all aspects" are also referred as metadata for the raw data. In ISIS, the capability of ICAT has also been extended to accommodate simulation data, also shown as a blue box in the Figure.

4.2. Infrastructure Components

As also shown in Figure 4, there are three types of infrastructure components:
- (Data repository) **clients** for managing the data in data repositories
- (Data repository) **utilities** for medicating the interactions between users and data repositories
- **Databases** (or data repositories) for storing data

4.2.1. Clients

The end users of our tools will be using the clients to manage data. In the context of our pilots, data means analysed or derived data generated along the analysis pipeline, *not* raw data gathered from instruments. Here, the word 'manage' can mean a wide range of operations upon data, including (but really not limited to):

Immediate needs of researchers:
- **Deposit** (also called ingest or archive)
- **Explore** (also called browse, e.g. navigate through a complex provenance chain of identified datasets for a series of experiments or simulations)
- **Restore** (This operation assumes researchers know what data they want to restore. It can be used together with the search or discovery operation.)

Medium term needs of researchers:
- Control (e.g. perform access control differentiating who can access what at what time for how long)
- Annotate (e.g. describe the context of the data ingestion, specific details of the data)
- Search and discover
- Organise

Advanced needs of researchers:
- Visualise (e.g. understand the meaning of data through a comprehensible format)
- Link (e.g. link up related datasets or investigations)
- Use, reuse and repurpose (e.g. use part of or a complete dataset to perform secondary analysis)
- Export (e.g. export to a different repository)

All these operations are important. Some are important because they fulfil the *immediate* needs of researchers. These include deposit, explore, and restore operations. As the requirement deliverable and our stakeholder analysis show, the solutions to these operations in the current data management landscape are fairly primitive.

Some are important because they fulfil the medium term needs of researchers. These include annotation, control, searching, discovering, and organising operations. However, without a proper solution to the immediate needs, the solutions to the medium term requirements cannot be satisfactory.

Some are also important because they allow advanced exploitation of data and offer maximum potential to the usage of data. These are visualisation, link, use/reuse/repurpose, and export. The advanced operations have to be built upon the solutions to the immediate and medium term needs.

Given the timescale of the project, it is clear that we should concentrate our efforts on building up infrastructure components to address the immediate needs, i.e. the components that allow deposit, explore, and restore data. Putting it plainly, these operations allow researchers to put (derived) data into repositories, to see what is in repositories, and to get back the data that they have previously deposit into repositories.

However, there is a profound meaning behind the operation of "putting data in" in the context of our project. What we are really capturing is data *and* the provenance about data (e.g. what program produces data, what program consumes data, what are the programs), rather than the data itself.

4.2.2. The Utility Components

These utility components facilitate the interactions between client tools and databases. In essence, they present the interface to data repositories. This interface refers to the *Application Programming Interfaces (APIs)* exposed by the services of data repositories. The extended CSMD model (i.e. the information model) underpins the services to guide data ingestion and restoration operations. The CSMD model is abstract so that it needs to be mapped into concrete data models for the implementation. Two types of data models are being developed: a XML schema representation and a database schema representation of the CSMD model. It should be pointed out that the model itself does not mandate the use of any specific relational database technologies (e.g. MySQL, Oracle).

To facilitate cross organisational data exchange, we envisage that all types of data repositories (personal, institutional, national, and international) implement the same information model.

All services manage data through the data access layer, which, in database terms, represents the persistent layer sitting on top of relational databases.

4.2.3. The Database Component

All data repositories are backed by databases, which implement the database schema defined by the information model.

4.3. The Pilots

The focus of the pilot is on the green boxes, each of which depicts an area where the (first) pilot implementation effort was spent and will be spent till the end of the pilot implementation phase (i.e. Dec. 2010). The detailed design and implementation of the

infrastructure components and the data models of the pilots will appear in the upcoming deliverables on Pilot Implementation and Information Model.

4.3.1. Principles for Developing the Pilots

The following principles are used while implementing the pilots.

Investigation of non-intrusive metadata capturing: Metadata is important. However, capturing metadata is often perceived as a labour intensive task that may not produce direct and immediate benefits to researchers. Thus, in reality, the task of recording and capturing metadata is often delegated to junior members in a research team. Hence, it is crucial to bear the non-intrusive requirement in mind when designing new tools or environment to support scientific research. This task investigates non-intrusive ways of capturing and recording metadata, aiming to reduce the amount of metadata capturing work to a minimum.

Progressive metadata management and maintenance: This task investigates ways of capturing metadata during the data analysis process of structural science work *as the analysed data is generated and collected by researchers*. This is *fundamentally different* from the conventional approach where metadata of research data is captured right before the data is ingested into institutional repository.

4.3.2. The Cross Organisation Pilot Implementation

This pilot investigates the data management issues arise from performing data analysis across organisational boundaries. As highlighted in the red box in Figure 5, this pilot focuses on the road map horizontally, embracing four levels of data repositories for scientists working across organisations: personal, institutional, national (service), and international (facility/service).
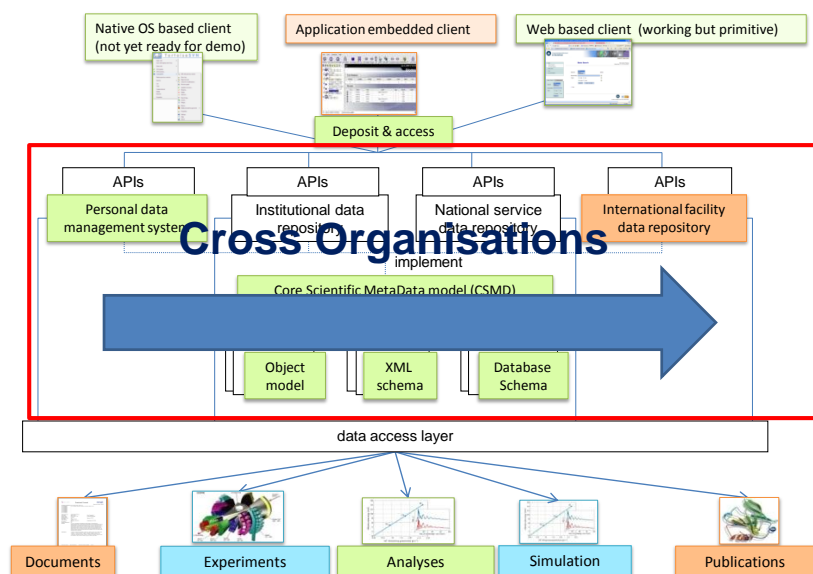
Figure 5. Illustrating the Cross Organisation Use Case in the Roadmap

4.3.3. The Cross Disciplinary Pilot Implementation

This pilot drills down to the details of the cross-disciplinary use case by applying the information model to the earth science data analysis pipeline. As highlighted in the red box in Figure 6, this pilot slices the roadmap vertically from the top to the bottom, addressing the data management needs of scientists *while they perform data analysis* after an experiment is completed. Compared with the cross organisational pilot, this specifically focuses the data management issues arising from the data analysis process. Such analysis can be performed for his own research (i.e. the personal scenario), collaboratively with others using institutional, national or international equipments or facilities (the institutional, national service, or international service scenarios).
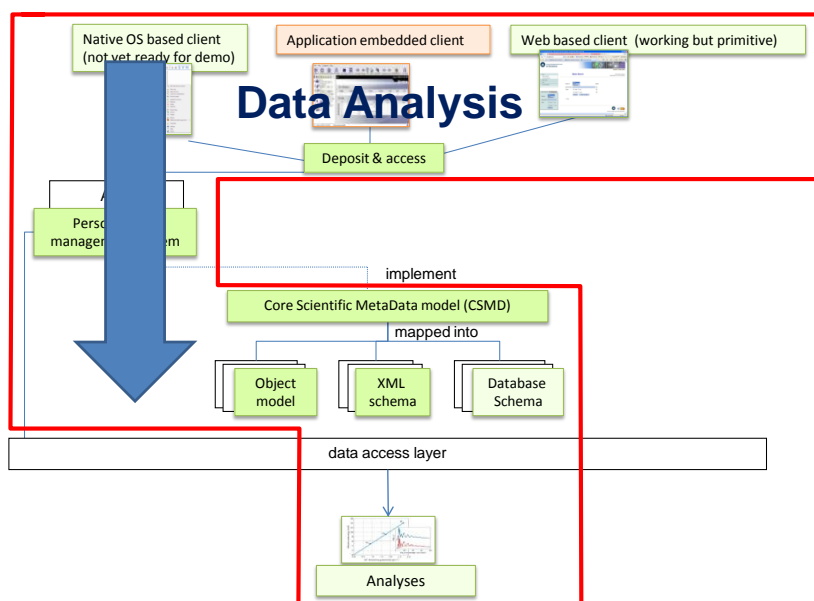
Figure 6. The Cross Disciplinary Pilot – the Data Analysis Scenario

5. Summary

This deliverable describes our rationales for choosing six out of the sixteen requirements identified in the requirement deliverable. These chosen requirements directly relate to the daily research data collection and analysis work that researchers perform. We hope addressing these issues would allow the improvement of researchers' productivities, and in the long run, leading to accelerated research discovery and investigation.

In addition, this report also outlines our plan and timeline for implementing the selected requirements as pilot implementations. As there has been a wealth of work in workflow management, we have specifically chosen a few areas which we believe will have significant impact on the success of the tools we develop.

References

[1] Manjula Patel, "*D1.1 Requirements Report*", Work Package 1, November 2009 – June 2010, JISC I2S2 project, July 2010, available at:
http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2-WP1-D1.1-RR-Final-100707.pdf

[2] Erica Yang, "*Martin Dove's RMC Workflow Diagram*", a supplementary requirement report, Work Package 1, November 2009 – June 2010, JISC I2S2 project, July 2010, available at:
http://www.ukoln.ac.uk/projects/I2S2/documents/ISIS%20RMC%20workflow.pdf

[3] Shoaib Sufi and Brian Matthews, "*A Metadata Model for the Discovery and Exploitation of Scientific Studies*". In Domenico Talia, Angelos Bilas and Marios D.

Dikaiakos (Eds.) Knowledge and Data Management in GRIDs, 2007, pp135-149, Springer: Berlin.

[4] Brian Matthews, Shoaib Sufi, Damian Flannery, Laurent Lerusse, Tom Griffin, Michael Gleaves, Kerstin Kleese. "*Using a Core Scientific Metadata Model in Large-Scale Facilities*". The 5th International Digital Curation Conference, London, England, 2-4 December 2009.

[5] Shoaib Sufi and Brian Matthews, "*A Metadata Model for the Discovery and Exploitation of Scientific Studies*". In Domenico Talia, Angelos Bilas and Marios D. Dikaiakos (Eds.) Knowledge and Data Management in GRIDs, 2007, pp135-149, Springer: Berlin.

[6] Erica Yang, "*The Pilot Implementation for Accommodating Derived Data in Scientific Data Analysis*", available at: https://www.jiscmail.ac.uk/cgi-bin/filearea.cgi?LMGT1=I2S2&a=get&f=/WP3-InfoModel-Pilots/ICAT-Lite-100614.ppt