Project Information			
Project Identifier	*To be completed by JISC*		
Project Title	Infrastructure for Integration in Structural Sciences (I2S2)		
Project Hashtag	i2s2		
Start Date	1st Oct 2009	**End Date**	31st March 2011
Lead Institution	Universities of Bath and Southampton		
Project Director	Liz Lyon (UKOLN, University of Bath)		
Project Manager	Manjula Patel (UKOLN, University of Bath)		
Contact email	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	Managing Research Data (Research Data Management Infrastructure)		
Programme Manager	Simon Hodson		

Document Information			
Author(s)	Neil Beagrie + Project Partners		
Project Role(s)	Project Partners		
Date	17th May 2011	**Filename**	I2S2-WP4-D4.1-CostBenefitsCastStudies-110517
URL	http://www.ukoln.ac.uk/projects/I2S2/		
Access	This report is for general dissemination		

THE I2S2 RESEARCHER AND SERVICE BENEFITS CASE STUDIES

Contents Pages

1. SYNTHESIS

1.1. BACKGROUND

The Infrastructure for Integration in Structural Sciences (I2S2) project is identifying requirements for a data-driven research infrastructure in ‘Structural Science’, focusing primarily on the domains of Chemistry and Crystallography. During the first phase of the project, we commissioned a comprehensive data management requirement report[1] for the structural science research arena. The major findings from the report were:

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

The implementation plan[2] for the I2S2 project was written after we 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. It narrowed down our efforts to a few key areas that need most attention. Specifically, we addressed six out of sixteen findings resulted from requirements gathering process namely:

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

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

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

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

  (2) Control files and parameters;

  (3) Versioning information;

  (4) Processing software;

  (5) Workflow for a particular analysis;

  (6) Derived and results data;

  (7) Links between all the datasets relating to a specific experiment or analysis.
- Changes should be easily incorporated into the scientist's current workflow and be as un-intrusive as possible;
- It was clear that the processing pipeline in many scientific experiments tend to be near digital, relying on suites of tools, applications software and very often customised software. There is therefore a need to document, maintain and curate such software and support its future development;
- The Core Scientific Metadata Model (CSMD) and its implementation in the ICAT database of the Science and Technology Facilities Council (STFC) is a good candidate for further development and extension to take account of the needs of organisations outside of the STFC.

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

1.2. THE BENEFITS CASE STUDIES

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

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

Each benefits case study has been able to draw on more detailed source material in two benefits use cases[3] prepared as cost/benefit deliverables for the project.

1.3. Benefits Identified

The primary or major benefits of implementing I2S2 identified by the two benefits case studies are:

- **Enhanced data management and long-term stewardship**. The immediate beneficiaries are the core research teams and their staff and close collaborators. The changes that take place as a result of the project will immediately impact on their working practices and the benefits to their research that follow (better science, higher productivity) will be felt quickly by these workers;
- **Rapid access to results and derived data**. There is a substantial anticipated reduction in the latency of information access for derived data or results data. At the present time, the way to obtain such data from one's colleagues is to ask, and typically the latency cost is of the order of one day to receive the data, which is borne by both the user and his/her colleague. Implementation of I2S2 can reduce a one-day latency of data access down to five minutes for these researchers;
- **Increased productivity through time savings and increased efficiency**. These are primarily appreciated by and visible at, the level of national facilities and services (or whole institutions) as economies of scale accumulate any time savings across multiple researchers, experiments and samples. The same benefits may be viewed as less significant or have lower impact at the level of individual researchers where this level of scaling does not apply;
- **Better and larger publication output**. The higher-education institutes, facilities and researchers will have a consequential benefit that accrues from a better and larger publication output;
- **Training**. New users will benefit enormously by having ready access to a wide range of well-documented data examples for tutorials and practice studies;
- **Software and tool development**. Code developers for the software and tools will benefit from having access to a wide and diverse range of well-documented data. There is a wide range of different use cases, and developers need access to a wide range of examples for testing purposes. Moreover, the number of use cases increases with time, and developers need to have access to an expanding range of examples and accompanying data;
- **Wider access and use**. Facilities will benefit by providing access to results and derived data as part of their services to researchers. There will be easier retrieval or revisiting of experiments long into the future. Other research teams will benefit from having access to this data for new analysis or comparative studies. This in turn will lead to a benefit for the scientific disciplines;

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

- **Reducing risk**. There will be less likelihood of introducing error into the safety or conduct of experiments as a result of better electronic information transfer and less manual transcription between systems;
- **Data publishing**. The ability of data to be fully validated and therefore openly published without further context (i.e. journal article) and an increased visibility of data with a secured longevity will mean increased citation and greater long-term effectiveness of the research;
- **Knowledge transfer**. There are companies that are now marketing lab-based x-ray sources optimised for obtaining PDF data. Researchers in Benefits Case Study 2 are collaborating with one, and for this company the benefits from I2S2 will be similar to those outlined above plus the ability to make demonstration data easily available. This is not merely good for one company's advertising; availability of lab-based equipment meets a real community need.

1.4. Potential Metrics Identified

Service productivity and efficiencies. I2S2 has developed an activity model of the scientific research data lifecycle and associated tasks[4]. Using this to structure analysis, the National Crystallography Service activities that are expected to be significantly changed and impacted by I2S2 are being benchmarked to allow "before" and "after" time measurements. These are being documented in I2S2 Benefits Use Case 1[5]. It should therefore be possible to calculate any work efficiencies and time savings after full implementation of I2S2. As noted above time savings and higher throughput are particularly significant benefits for services because of economies of scale effects they can have dealing with many individual researchers, samples, and experiments. At the same time, metrics for these benefits are particularly difficult to capture within the timeframe of short projects or the limitations of pilot implementations but benchmarks for longer-term evaluation can be established.

Extending, training, and self-starting the user community. In Benefits Case Study 2, the wider user community is currently relatively small, producing about 15 papers a year. However, it is anticipated to grow extremely rapidly in the UK, in part promoted through the availability of new instruments at ISIS and Diamond. This leads to important requirements for training and ongoing code development, for which availability of well-documented data is a key requirement. The number of users and completed studies can be counted through the number of publications that cite the main program publication (Journal of Physics: Condensed Matter 19, 335218, 2007). As of the time of writing, citations number 38 (15 a year for the past two years; we include self-citations here because many of the self-citations are in collaboration with new users, which is a typical trend for any method when new users need help). A clear metric of success here will be an increase in the citation rate of this paper.

New users based on ISIS are inevitably going to grow slowly, because beam time on ISIS instruments is limited. On the other hand, there are potential gains in the user base to be found in the use of new neutron facilities (e.g. at reactor sources on specialised instruments,

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

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

and on existing and new spallation sources) and also in the use of x-ray scattering methods at synchrotron sites and new laboratory sources. Our base of data here is rather limited but could easily be expanded. We will be able to track expansion of the use of our methods to these new instruments/sources from the Science Citation Index to give a clear measure of success.

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

Improved software and tools. In Benefits Case Study 2, the benefits for code development come through having a wide range of available examples. For example, we need ready access to data for magnetic materials, non-crystalline materials, and materials containing molecules, for both neutron and x-ray data. The two markers of success are a) as we develop new functionality we can turn around a suite of test data; b) that as new types of systems demand new functionality, we have the ability to add new data sets to our test suite.

1.5. Conclusions

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

The identified benefits can be divided into two major areas:

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

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

2. I2S2 Case Study 1 (Service Perspective): National Crystallography Service

2.1. Background

This I2S2 case study explores the data management interactions between a 3-tiered chemical crystallography system that we operate in the UK – the home laboratory or crystallography facility (as exemplified by the University of Cambridge), a mid-range facility (UK National Crystallography Service based at the University of Southampton) and a centralised facility (Diamond Light Source Ltd). We are primarily concerned with addressing two different aspects: cross-institutional data management and improving the working efficiency of the researcher. The scenario incorporates three different institutions, with different working, administrative and data management practices in structural science.

In the first instance, a scientist from a small research group needs to get a crystal structure run on her new compound and therefore submits it to her departmental crystallographer. An experiment shows the crystals to be too small to obtain a usable dataset on local instrumentation and therefore the crystallographer submits it to the National Crystallography Service (NCS). Examination at the NCS yields a dataset, but the workup gives a result that is not of a publishable quality. The NCS then schedules an examination at the synchrotron central facility – Diamond. Data collection at Diamond provides a publishable result, which is communicated back to the originating researcher and deposited in the NCS repository alongside all administrative and supporting data/information gathered throughout the entire process.

2.2. Established Practice and Challenges

The key aspects and steps of the crystallography experiment are essentially the same irrespective of the facility at which it is being performed. These are namely crystal selection, data collection and processing, structure solution and refinement, structure analysis and publication. Until recently (10 years ago) this process would have been performed by the single academic researcher in the department who has an interest in this technique and also the amount of data generated, due generally to a relatively slow speed of the instrument, was considered to be manageable by this 'lone researcher'. The last decade has seen an explosion of data and an increase in the use of the technique due to the speed and accuracy now possible. Accordingly some laboratories have increased in size, both in terms of the number of researchers and number of machines and also this technique has been made available at high-powered centralised facilities.

Data management is becoming a very serious issue and one which most researchers in this field have now realised as a pressing problem. The home laboratory has informal methods for interacting with its users and little established protocol for data management and preservation, however centralised facilities have to control access very closely – often by application and there is an onus on the provider to properly curate the data that they are collecting for or on behalf of others.

2.2.1. Data Exchange and Transfer

The chemical crystallography community has adopted and been using a common format for exchange and publishing of results data for over 15 years and is a leading example of the success of this approach. The crystallography community has a long (40 years) established centralised database which has been mining the data from publications and making it available as a collection – this process is now embedded in the publication of research work as data must be deposited with the database prior to submission to a journal.

However, certainly for larger facilities, there is a requirement to be able to handle and manage all the data relating to a particular experiment, i.e. application, safety, information relating to the experiment / process, raw data and all derived data leading to the result. This is generally rather poorly done outside of central facilities, with the primary focus being on the result – however often this information is important when transferring the sample / experiment between institutions and is vital when curating an experiment so that it is fully understood when returning to it some considerable time in the future.

2.2.2. Institutional Working Practices

Established working practices are subtly different between small-scale institutions. Moreover, most crystallography labs have been in existence for 20-30 years and are therefore not particularly integrated into any digital management infrastructure. These working practices do not map easily on to the formalized approach adopted by mid-range and centralized facilities.

2.2.3. Paper-based Documentation and Limitations of Electronic Forms for Chemistry

Laboratory notebooks are still paper-based and virtually all crystallography services capture the requisite information on paper forms. However, the data is captured and worked up in electronic form so there is a format mismatch between the data itself and information relating to it. There is a distinct reluctance of NCS users to submit samples via electronic forms (this has been tried in the past but there is difficulty with generating chemical schemes in electronic submission systems - these are possible, but take considerably longer to do than drawing on paper!). The NCS currently obtains all the vital information regarding samples from its users on paper forms.

It has long been necessary for synchrotrons and central facilities to consider data management, however the systems that have been developed are heavyweight and also do not transcend the boundaries of the facility.

2.2.4. Different Individual Researchers' Approaches to Data Management

Finally, individual researchers have their own approaches to data management – virtually all are fundamentally flawed. It often takes considerable time and effort to find old datasets (a problem that gets worse with increasing scale, especially in a lab like the NCS which collects >2000 per annum). Moreover the administrative requirements (sample logging, safety information, access, etc) of many of these institutions are significant and often manually managed. These factors considerably hamper researcher efficiency.

2.3. Benefits from Project

The I2S2 project will address the challenges highlighted above in the following ways:

2.3.1. Data Exchange and Transfer

After the implementation of the I2S2 management framework, metadata, information and data itself will flow seamlessly between institutions. The administrative and associated information regarding a sample or experiment will be set in a descriptive framework that can sit alongside the formalised approach that the community already use. Within this framework the data and associated metadata can be integrated for management purposes, which would result in a simpler and seamless way to give data context, associate safety and administrative information and be consistent with the way in which larger scale operations are conducted.

2.3.2. Institutional Working Practices

The framework that I2S2 is developing is able to describe activities and therefore capture the relevant metadata associated with a particular process independently of how it is conducted. Different scientists operating under subtly different working practices will be able to manage their data and associated information under the same framework and thereby benefit from all the advantages that a common standard provides.

2.3.3. Paper-based Documentation and Limitations of Electronic Forms for Chemistry

The I2S2 project will map the information obtained through submission forms into the framework and thus enable the development of electronic infrastructure that can be used alongside the paper-based system. Users of a facility may then chose to use either method, but in the case of the paper-based approach NCS administrators can easily transcribe the information into the correct form. As far as NCS is concerned, some electronic submission is better than the current situation of none and at least the administrators will have a simple mapping between the two systems which will make the transcribing job easier and less error prone. Recent software developments in the computing science and chemistry communities will lower the barrier to adopting electronic sample submission: society in general has become considerably more familiar with interacting with the digital world in the last 2-3 years; access to such systems is less torturous than a few years ago; new chemical structure drawing applets make it considerably easier to sketch online (although not as rapidly as actual drawing); electronic systems can now provide something back to the submitter, whereas before the information flow was very much one-way (eg telemetry and tracking, remote experiment monitoring, data interaction and visualisation).

2.3.4. Different Individual Researchers' Approaches to Data Management

Results data are predominantly stored on the file system of the individual researchers personal computer. Accordingly there are little or no metadata and identifiers, except the name of the folder and the contents of its constituent files. Different stages of an analysis, or iterations of the same stage e.g. trying different approaches, are either performed in different sub-folders or by using different file name prefixes. There is a lack of convention for naming folders and sub-folders – definitely globally and often locally. This means that there are almost as many ways of organising data as there are researchers out there!

Archives are traditionally 'back-ups' i.e. an image of the file system on which files are stored and contain no metadata other than the filename. Back-ups are predominantly made on removable media (i.e. floppy disk, pen drive, CD, DVD), but with the increasing size of hard disks in personal computers the tendency is to store everything on the laptop and back-up the whole data directory. Crystallographers have a tradition of backing up raw data. This is still predominantly on DVDs or removable solid-state media, although we are beginning to see a very gradual emergence of on-line storage – mainly spearheaded by national facilities.

2.4. Summary and Key Points

In performing this case study we are probing cross-institutional and scale aspects of I2S2. It is often possible to 'manage your own affairs' within your institution, but migrating across administrative boundaries is often very difficult and time consuming to do.

The major challenges facing I2S2 are therefore achieving widespread adoption and seamless use of mid-range and centralised facilities, changing crystallographers mindsets to be able to consider separate administrative and scientific entities as equally important parts of the archived record, enabling the scaling up (or perhaps more importantly, down) of the I2S2 methodology, and finally to change the perceived opinion that this is all about 'open access' which immediately puts many chemists on their guard.

The primary or major benefits of implementing I2S2 are:

- More time for researcher to conduct their work by saving time and thereby increasing the efficiency of work in:
    - Writing proposals
    - Accessing facilities
    - Writing reports and papers
    - Backing up results
    - Auditing outputs
- Increased facility productivity through increased administrative efficiency;
- Less likelihood of introducing error into the safety or conducting of experiments;
- Easy retrieval or revisiting experiments long into the future;
- Ability of data to be fully validated and therefore openly published without further context (i.e. journal article):
- An increased visibility of data with a secured longevity will mean increased citation and long-term effectiveness of research.

Service tasks which are expected to be significantly changed and impacted by I2S2 are being benchmarked to allow "before" and "after" measurements. These are being documented in the relevant I2S2 Benefits Use Case (Use Case 1). It should therefore be possible to calculate any work efficiencies and time savings after implementation of the pilot in I2S2.

3. I2S2 Case Study 2 (Researcher Perspective): Reverse Monte Carlo studies using neutron total scattering data

3.1. Background

Many phenomena in materials physics and chemistry are partly or wholly understood in terms of how the atoms behave over very short distances, and often it is important to understand fluctuations in local atomic structure compared to average structure. One experimental probe of local structure is total scattering of neutrons or x-rays, which is essentially measuring the entire scattered diffraction pattern, including both Bragg peaks and the background scattering. The Fourier transform of the total scattering function is the pair distribution function (PDF), which in practice is a histogram of instantaneous interatomic distances weighted by the scattering power and concentration of each atomic species. Because the scattering power depends on the type of radiation, neutron and x-ray beams are complementary, although from our perspective the art is better developed for neutron beams.

Whilst the PDF contains useful information about local atomic structure, it can be hard to interpret unless one does so in the context of model building. We use a Monte Carlo method – called the Reverse Monte Carlo (RMC) method – to construct and refine atomic configurations, and the end result is a set of configurations whose calculated PDF and scattering function is in best agreement with data. The task of interpreting the PDF is transferred to the analysis of these configurations. We believe that we have demonstrated that this approach yields very reliable information as judged by a number of different measures, and often this information cannot be gleaned from other approaches.

The data challenge comes from the fact that there is quite a complex workflow going from the raw data to the final configurations, and since the quality of the raw data is high it is essential that the quality is preserved at all stages in the workflow. This workflow involves transforming and merging the raw data into usable derived data (diffraction profiles, experimental PDF etc), performing Rietveld refinement on the diffraction part of the data to extract a starting structure and instrument parameters, performing the RMC, and finally running the analysis of the configurations (perhaps using bespoke tools). Each stage will generate its own set of files, which are used to derive input data for other tasks (for detailed discussion of the Earth Sciences Cambridge and ISIS workflow see the I2S2 Main Report (http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2-WP1-D1.1-RR-Final-100707.pdf).

3.2. Established Practice and Challenges

The team that we are concerned with here consists of Prof Martin Dove (Cambridge, I2S2 co-investigator) and collaborators at the STFC ISIS facility and University of Oxford. The wider teams are of two types. First are the PhD students and research workers, mostly in Cambridge and Oxford. In this case people are part of the team for shorter periods of time than the useful lifetime of their data. The second type are other research groups who are working with us to learn and develop their own expertise in the methods. This group is important to note, because part of the challenge with data is to have good data sets for new users to train with.

The wider user community is currently relatively small, producing about 15 papers a year. It is anticipated to grow extremely rapidly in the UK, in part promoted through the availability of new instruments at ISIS and Diamond. This leads to important requirements for training and ongoing code development, for which availability of data is a key requirement.

3.2.1. Data management and long-term stewardship

The key challenges for the whole programme of work within this case study are to turn short-term personal-space based management of data on individual PCs into long-term stewardship of data in a managed repository, with the focus on the data derived by the user, notably the corrected data that are used in the Rietveld and RMC stages, and the resultant configurations and results of analysis.

The challenge here is to be able to upload to a permanent repository the outputs from each stage of data analysis with useful metadata annotations. The various stages include the production of the derived data that forms input into the data analysis, the Rietveld refinements, the RMC outputs, and the detailed analysis. There is also the need to capture the workflow, and also bespoke analysis tools (whether spreadsheets, scripts for analysis tools such as Matlab, or bespoke programs).

3.2.2. Extending the user community: training and self-starting

Our long-term goal is to extend the user community. At the present time, new users work with the team, attend training courses, and have a small set of on-line tutorials. We recognise that these are not extensive, and one clear challenge is to be able to turn completed studies into training modules, recognising that there are many different scientific use cases that are not adequately reflected in a small number of examples.

3.2.3. Higher work throughput and outputs through access to data

At the present time, the way to obtain data from one’s colleagues is to ask, and typically the latency cost is of order one day to receive the data, which is born by both the user and his/her colleague. There are several cases where ready access to managed data would make an identifiable change:

- When users pick up work after a break, and need to spend time ensuring that they are working with the best set of data files. This includes when a new worker picks up someone else’s work to continue with, for example a supervisor trying to write a paper based on the work of a recent PhD student;
- For studies that build on previous work;
- For ongoing code development by the team, when access to a wide range of data examples is needed for testing purposes. As in training, there is a wide range of different use cases, and developers need access to these data. Moreover, the number of use cases increases with time, and developers need to have access to an expanding range of examples.

3.2.4. Sharing outputs

When we give talks or prepare review papers, we need to present a wide range of use cases. As in the previous challenge, we typically have to ask our colleagues for data, with the inevitable latency.

3.3. Benefits from Project

3.3.1. Data management and long-term stewardship

The detailed benefits and beneficiaries are outlined in the subheadings below; here we will focus on some of the key indicators of success.

- The immediate beneficiaries are our core team and their immediate staff and close collaborators. The changes that take place as a result of the project will immediately impact on their working practices, and the benefits to their research that follow (better science, higher productivity) will immediately be felt by these workers.
- The higher-education institutes and facilities will have a following benefit that accrues from a better and large publication output.
- New users will benefit enormously by having ready access to a wide range of examples for tutorials and practice studies.
- Code developers will similarly benefit from having access to a wide and diverse range of data.
- Facilities will benefit by providing access to data as part of its service.
- Other research teams will benefit from having access to data for new analysis or comparative studies. This in turn will lead to a benefit for the scientific disciplines.
- There are companies that are now marketing lab-based x-ray sources optimised for obtaining PDF data. I am developing a collaboration with one, and for this company the benefits from I2S2 will be similar to those outlined above plus the ability to make demonstration data easily available. This is not merely good for one company's advertising; availability of lab-based equipment meets a real community need.

Some of the benchmarks are described in the following sections.

3.3.2. Extending the user community: training and self-starting

The beneficiaries are new users, who learn how to use our tools. The number of users and completed studies can be counted through the number of publications that cite our main program publication (Journal of Physics: Condensed Matter 19, 335218, 2007). As of the time of writing, citations number 38 (15 a year for the past two years; we include self-citations here because many of the self-citations are in collaboration with new users, which is a typical trend for any method when new users need help). A clear metric of success here will be an increase in the citation rate of this paper.

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

3.3.3. Higher work throughput and outputs through access to data

The beneficiaries that will provide the benchmark here are my immediate team of colleagues and research staff (across Cambridge, STFC and Oxford). The indicator of success is that we can turn a one-day latency of data access down to five minutes.

The benefits for code development come through having a wide range of available examples. For example, we need ready access to data for magnetic materials, non-crystalline materials, and materials containing molecules, for both neutron and x-ray data. The two markers of success are a) as we develop new functionality we can turn around a suite of test data; b) that as new types of systems demand new functionality, we have the ability to add new data sets to our test suite.

The benefits for collaborating scientists come from having access to each other's data. One example is the ability to access data of a PhD student after they have left in order to write papers. I have cases at hand that will highlight success. I also have two new PhD students who will start to carry out this work, and if we have processes in place the availability of output of her work will demonstrate success; the same is true for my colleague in Oxford.

3.3. 4. Sharing outputs

We all give talks, whether in conferences, lectures or training courses. In many of these cases we need a wide range of data to be able to explain the methods in a comprehensive manner and to give added interest. Often preparing presentations or lectures takes place against a hard deadline, and often compromises are made because of the lack of access to data. The benefits to the speaker are clear in giving access to all the examples possible. Similar to the above point, we are aiming to turn a one-day latency down to five minutes.

3.4. Summary and Key Points

The key point from this use case is that with a multi-component workflow that generates and uses many files, we need to change the way we work to ensure the safe long-term preservation of data in a way that allows for easy sharing of files.

The key groups who will benefit from the project are a) immediate research teams (for their science and for training purposes), b) new users, c) code developers, d) facilities, e) the scientific disciplines, and e) industry.

The benefits for immediate research teams are access to completed data and access to data in progress to enable re-use for publication, presentations and subsequent follow-on analysis.

The benefits for new users are a wide selection of tutorial examples.

The benefits for code developers are a wide selection of test cases to enable new functionalities to be thoroughly tested.