Project Information			
Project Acronym	I2S2		
Project Title	Infrastructure for Integration in Structural Sciences		
Start Date	1st Oct 2009	**End Date**	31st March 2011
Lead Institution	Universities of Bath and Southampton		
Project Director	Liz Lyon (UKOLN)		
Project Manager	Manjula Patel (contact details: 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	Managing Research Data (Research Data Management Infrastructure)		
Programme Manager	Simon Hodson		

Document Name			
Document Title	I2S2 Benefits Use Cases		
Reporting Period	*N/A*		
Author(s) & project role	Neil Beagrie (Charles Beagrie ltd),Simon Coles (University of Southampton), Martin Dove (University of Cambridge)		
Date	30/03/2011	**Filename**	I2S2_BenefitsUseCases_final
URL			
Access			General Dissemination

Document History		
Version	**Date**	**Comments**
0.1	22/02/2011	Draft version of use cases and introduction for review
0.2	30/03/2011	Final version

I2S2 Benefit Use Cases

Authors: Neil Beagrie (Charles Beagrie Ltd), Simon Coles (University of Southampton), Martin Dove (University of Cambridge)

Contents

1. Introduction

The Infrastructure for Integration in Structural Sciences (I2S2) is identifying requirements for a data-driven research infrastructure in 'Structural Science', focusing primarily on the domains of Chemistry and Crystallography. A key aim of I2S2 has been to develop use cases that examine the business processes of research, identify the costs and benefits of the integrated approach proposed by I2S2, and explore perspectives of "scale and complexity" and "research discipline" throughout the data lifecycle. During the course of the project, the complementary but often different perspectives of researchers and central facilities in terms of benefits were also recognised as significant and built into the use cases. The I2S2 Cost/Benefit Use Case 1 (National Crystallography Service) 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 Cost/Benefit Use Case 2 (Prof Martin Dove, University of Cambridge) applies the approach to Mineral Sciences and interactions between individuals, collaborative research groups and facilities, and provides a researcher's perspective of benefits.

Both use cases examine the business processes of research, and the benefits of the integrated approach proposed by I2S2. Each forms the source material for the Benefits Case Studies contributed by the project to the Managing Research Data Programme Benefits Synthesis Report.

1.1 Methodology

The approach and template for the value-chain and Impact analysis tool was developed by Neil Beagrie of Charles Beagrie Limited who acted as advisor on cost/benefit analysis to each of the domain experts (Simon Coles and Martin Dove for Use Cases 1 and 2 respectively). The domain experts drafted additional text and commentary for each of the elements in the use cases with further editorial feedback and guidance provided by Neil Beagrie. Drafts were then circulated for comments to the wider I2S2 project team.

The template drew on the following elements to populate it:

Activity – This is derived from the activities in the I2S2 Idealised Scientific Research Data Lifecycle Model. Domain experts can omit any activities not relevant to their use case.

Change in and Benefit from I2S2 – A generic initial element set of potential benefits from KRDS was provided for the template. The domain experts identified changes in I2S2 and added to and modified benefits as needed.

Impact Weighting – An impact weighting was assigned by the domain expert (from 1 [low] to 5 [high]). This was done to aid selection of benefits/changes with greatest impact for further analysis. This is a subjective weighting reflecting specific perspectives of benefits to different parties (i.e. impact weightings for the same activity can vary between the use cases). Activities may not be of equal scale (e.g. archive activities are conflated in the I2S2 model and not sub-divided compared to the level of definition of researcher activity) and alongside potential and feasibility this can be reflected in impact weightings.

From When? – The domain expert entered a specific year (Year 1, Year 2, Year 3, etc) from when the benefit should be realised.

Who Benefits? – The domain expert identified the beneficiaries (researcher, research group, institution, research funder, discipline, etc) of each benefit.

Key Resources – A pre-defined list agreed in consultation consisting of time, salaries, full economic costs, facility costs, number of samples processed/data sets collected, number of users of the facility, and percentage of Users/Community who will benefit.

Measurable Impacts and Qualitative Impacts – elements discussed and agreed individually for each use case.

Selection for Further Analysis – as time and resources precluded detailed examination impact of every benefit identified, the impact weighting was used to select any changes/benefits for further analysis. For I2S2 Cost/Benefit Use Case 1 (National Crystallography Service) changes in five activities combining high impact and measurable or identifiable impacts were selected for further in-depth quantitative costing analysis. There are a number of caveats that should be considered when undertaking such an operation:

1. A consideration of the confidence in measuring key variables – this could vary considerably and will affect the assessment of the scale and impact of benefits. The impact weighting could vary by as much as 1 in either direction, depending on who is performing the assessment and under which regime/laboratory it is being performed. This is being mitigated by the same people performing the measurements in the same manner both before and after the implementation of I2S2.
2. What is the cost of defining variables and impacts? The experience of this work is that a considerable time element is involved in understanding how to measure these impacts, at what level of granularity to measure and then actually measuring them. The I2S2

project has funded a large component of the resource required to address this novel assessment – initial findings will form a useful set of guidelines for others.

3. How many times should a measurement be made in order to get a statistically meaningful average? Small repetitive measurements that are many in number have been averaged over numerous exercises eg ‘logging in’ samples occurs several times a day, every day and averages were measured over two separate week-long periods. One-off tasks were assessed for their ‘standard’ nature and appropriate adjustment made if judged necessary – these types of task can only be measured once every month – year.
4. Timescale of the Project and I2S2 Implementation. “Before” and “after” measurements cannot be undertaken in the lifetime of the project. “Before” measurements have therefore been made as a benchmark against which impact can be measured during piloting and full implementation at a later stage.

2. Use Case 1: A Service Perspective

This use case utilises the I2S2 lifecycle activity model as a starting point for analysing the changes and benefits arising from I2S2 and will feed into a second stage I2S2 benefits case study. It is written from a service perspective (Simon Coles, National Crystallography Service).

2.1 Value Chain and Impact Analysis

Activity	***Change in and Benefit from I2S2***	***Impact Weighting*** *(1 high- 5 low)*	***From When?*** *(year1-5+ etc)*	***Who Benefits?*** *(researcher, research group, institution, researcher funder, discipline, other)*
Research Concept	Data readily available. Easier hypothesis testing.	4	Year 3	Researcher
Write Proposal	Availability of data for track record / justification; Availability of administrative information from e.g. previous proposals, profile, etc so no re-keying.	1	Year 2	Researcher
Peer Review Proposal	All data easily available allows faster and less subjective or ambiguous review.	4	Year 3	Funder
Start Project	User and project information from proposal already available so some time saving.	3	Year 3	Research Group
Acquire Sample & Collate Information	Information supplied by user in form that can be propagated through the system; ERA automatically completed so no re-keying.	1	Year 1	Facility
Conduct Experiment	Sample information can easily be assigned to an experiment allowing easy management of both data and supporting information together	3	Year 1	Researcher
Process Raw Data	Full sample metadata record available. Available for preservation & curation and researchers to revisit.	2	Year 1	Researcher
Analyse Derived Data	Full sample metadata record available. Available for preservation & curation and researchers to revisit.	2	Year 2	Researcher
Interpret Results Data	Full sample metadata record available. Available for preservation & curation and researchers to revisit.	2	Year 1	Researcher
Documentation,	Majority of metadata already captured and automatically	1	Year 1	Researcher,

Metadata and Storage	assigned, so no re-keying.			Facility
Archive, Preservation and Curation	Committing to archive seamless and immediate so time saved and preservation assured.	2	Year 1	Facility, Institution
IPR, Embargo and Access	Management of embargo automatic so no time spent monitoring or managing public release.	2	Year 2	Facility, Institution
Write Usage Reports	Full data immediately available so no time wasted in searching and presenting, leaving time to discuss results.	1	Year 1	Facility, Funder
Submit to Publications Database	Immediate and seamless deposition / harvesting so one less job to do.	3	Year 2	Researcher, Funder
Prepare Supplementary Data	Full data record available in structured form potentially enabling automatic preparation.	1	Year 2	Researcher, Facility, Community, Publishers
Prepare Manuscript	Supporting data readily available so don't have to find, format and present.	2	Year 2	Researcher, Facility
Peer Review Research	Full data record & provenance available for checking and validation.	3	Year 3	Funder, Publisher, Community
Publish Research	Automatic; Increased visibility and/or citation of data.	2	Year 2	Researcher, Publisher, Facility, Institution, Community
Validate & Re-use (a) by original team (b) by others	Full data record & provenance available for re-use in new experiments; Full data and provenance available for preparing lectures, talks, and review papers; Full data and provenance available for generating new tools and developing existing programs; Full data and provenance available for training of other users, including making available extensive suites of data that users can select as best match their own interests;	1	Year 4	Researcher, Community

	Published datasets are complete and accompanied by full metadata / documentation, which allows rapid assessment of validity, correctness and reusability; Immediate and seamless incorporation into other bodies of data.			

Key Resources

The key resources that are quantifiable in order to measure the scale and impact of benefits arising from the I2S2 implementation are:

1. Time spent on an activity or facility: This is a measurable factor and can easily be recorded before and after I2S2 changes – it is expected that this will have the most impact on the act of performing research by saving researcher and administrator time. All research-based tasks will take differing amounts of time, however they can broadly be classified into different categories and therefore the most sensible approach is to measure a number of these and take an average. Administrator tasks are more formulaic and repetitive, however it is still better practice to average over a number of measurements.
2. Staff Salaries: The financial benefits, mainly due to time savings, as measured by salaries are enormous and the most impacting. Costings have been performed for Principal Investigator, Administrator, Post-Doctoral Research Assistant and PhD Student and should be considered as including NI and pension costs. Average pay scales for each role should be used.
3. FEC on salaries: This rate varies from institution to institution and therefore an average of 80% is assumed.
4. RAL Facility Cost: This is a very difficult cost to estimate – several different costings may be made depending on what level is being looked at. Ideally the cost for Diamond and ISIS per day/hour and what this includes should be considered.
5. Number of Users of Facility: With a streamlined access and data management infrastructure it is likely that a facility could enable access for more users. This is a clear benefit where the cost would most appropriately be linked to the facility operation / provision cost.
6. Number of samples processed / data sets collected: This is a clear benefit and can be measured easily, but costing is slightly more difficult to assess.
7. Percentage of Users/Community who will benefit from I2S2 changes: In some cases the entire community would benefit and in others a lesser grouping.

Measurable Impacts

For the following important impacts from I2S2 it is possible to quantify and devise metrics for:

1. Time savings for researchers and support staff; the impact of good data management practice and supporting infrastructure will be greatest on staff time. For short, simple, repetitive and formulaic tasks an actual clock time is recorded and for more complicated and infrequent tasks an estimate based on a single process is made.

2. Time savings on facility; this impact is again readily measurable and likely to be significant – it is based on the summation of all the staff time savings and all the samples processed.
3. Increased output; gauged by the number of samples and/or datasets processed. This impact will not be easily noticed or measured over a short timescale e.g. a day, but is likely to be significant over longer timescales e.g. a year.
4. Percentage of research data from experiment made available for re-use; This measure is currently the number of structures published in journal articles (some 200 or so out of 1652 collected every year) – with proper data management and control this amount could in theory approach 100% (although it never will as some data are considered to be commercially sensitive and would never be made publicly available).

Qualitative Impacts

Other important impacts we believe I2S2 will deliver but will be difficult to measure due to timescales or other factors include:

1. Quality of research; it is not possible to measure this in terms of the number of outputs, however thorough data management will ensure that the data itself is sufficiently self-describing to enable it to be published, curated and reused in its own right without the necessity of a formal journal article. The provenance information surrounding the creation of a piece of data is thereby available and therefore data can be validated long into the future.
2. Citation of data; with data properly structured it is possible to provide a long-term framework to support it – this provides the means and confidence for referencing it and for that reference to be persistent. Additionally data can be cited without being part of a formal journal article and therefore more citation may be made and they can be made directly to the data as opposed to an article encompassing it.
3. New research opportunities; currently most research ideas and opportunities are made based on the published literature – this makes it difficult to find the actual data and also increases the timescales for making the data available. More data can be made available and immediately discoverable.
4. Validation of research; Full provenance information collected throughout the experimental process and at the point when it was conducted will be available – this enables a rigorous assessment of the quality and validity of the data both at the time it was generated and also long into the future.
5. Knowledge transfer to industry; there is currently a significant divide between academia and industry which could be bridged somewhat if data from research are made available in a trusted and structured form for commerce to evaluate, develop and take to market.

Selection of Activities by Impact for Further Analysis

The activities assessed as having high impact from changes in I2S2 have been selected for further analysis including quantification of benefits wherever possible as follows:

1) Write proposal

There are a number of different proposals that could be considered here (research council, institution, charity, knowledge transfer, central facility access, etc), however the most appropriate and impacting for I2S2 would be for access to central facilities and therefore measurements are made on this basis.

This data is based on the time taken for a PI to write an application (based on SJC Diamond application Sept 2010) and would normally occur twice a year in the case of applying for beam time (however other applications e.g. to research councils, would be made in a more sporadic fashion).

Task	*Before I2S2*	*After I2S2*
Work up and write concept and provide supporting data	2 hours	to be determined
Include track record / previous usage outcomes	15 minutes	
Include publications list	15 minutes	
Prepare safety and sample information	15 minutes	

2) Acquire Sample

The primary factor here is administrator time and this has been averaged over a two-week period – the NCS is contracted to process 1652 samples per annum. Timings are provided per sample for:

Task	*Before I2S2*	*After I2S2*
Logging in	2 minutes	to be determined
Logging/sending out	6 minutes	
Adding to the Experimental Risk Assessment	3 minutes	

Caveats:

i) In order to obtain accurate timings the tasks had to be rigidly defined. There are numerous other tasks that are performed under the 'acquire sample' category that could have been added to this process. The primary task here is around recording information for reporting facility usage – the funder requirements change quite regularly in this respect and if the system cannot record the appropriate information, then this has to be performed 'manually'. For example the recording of the new statistics on attempts, sample classification, outcome, adjusting of allocation based

on outcome and priority have not been recorded in the past and therefore there is no mechanism within the current system for doing so. Currently this additional data is recorded in a separate spreadsheet and amalgamated later and this adds 4-5 minutes per sample.

ii) Logging in is faster when you have a large number of samples from the same user.

iii) If logging out is done in a smaller number of large sessions, as opposed to regular short sessions more samples can be included in a single packet to a particular user.

The qualitative benefits are around safety information – where the right risk assessment data can be provided by sample originators and properly and accurately propagated through the system. This has a potential saving in terms of central facility costs if this regulatory data is trivial to import, manipulate and process. Additionally the information generated at this stage can be communicated to a user in a timely fashion and also used in drawing up facility service provision reports.

3) Documentation, Metadata and Storage

Currently this work is performed separately for both raw and results data and having devoted a lot of research funds (many different sources) into the infrastructure management, the major factor contributing to this activity is researcher time spent. Much of the raw data is organized, formatted and moved by automated scripts and this would have to occur for all 1652 samples processed each year. Average timings per dataset are:

Task	*Before I2S2*	*After I2S2*
Tidying up	2 minutes	to be determined
Generate metadata	1 minute	
Deposit	1 minute	
Include in spreadsheet record	30 seconds	

Results data requires more care and attention, as it is generated in a much less constrained (individuals personal computers) and structured (numerous different software packages) environment. This work is done by a researcher and would be performed on approximately 400 samples per year. Average timings per dataset are:

Task	*Before I2S2*	*After I2S2*
Tidying up	10 minutes	to be determined
Deposit	5 minutes	

In other laboratories this work would take a considerably longer time, as NCS has the appropriate policy and infrastructure in terms of personnel to do this work. Moreover, the NCS has been an innovator in this area and so has attracted research funds to develop systems to

support this type of work. In total it has taken about 3 months PDRA time to develop support for raw data and 3 years PDRA time to develop eCrystals (the repository system for crystal structure data which in part fulfils this role).

4) Prepare Usage Reports

The most significant factors associated with preparing usage reports are:

- Collating the information relating to submissions/number of experiments e.g. when, where, how many, how long, success, fail, etc for a period
- Collating the outputs information required eg degree of success (publishable, OK, some useful information, not really usable), nature of output (repository or database record, progress report, thesis, poster, journal article, press release or special interest magazine).

The length of time spent on this activity can vary dramatically depending on its purpose e.g. NCS usage (facility user), beam time usage (PI), grant award final report (PI), PhD student progress reports (PhD student) and thesis (PhD student).

For the purpose of this work NCS usage is the most appropriate example to consider. In surveying some select users the response is uniformly of the nature “it takes me about half an hour” and the tasks involved are:

- Working out which samples were submitted and looked at during the period in question.
- Tracking down the outcome of experiments e.g. it didn’t work; it gave some results that we could learn from but not use; it gave some results we can use; it gave a satisfactory result; it gave an excellent and easily publishable result.
- Finding and importing references to published work, conference talks or posters, theses, etc.

NCS currently requires the submission of very traditional usage reports (text and references in an electronic document) twice a year, however as part of the I2S2 implementation it is intended that this activity is completely overhauled and most of the required data will be immediately available. There is a direct, linear, relationship between the number of samples submitted by a user and the time taken to generate a usage report.

5) Prepare Supplementary Data

This activity is always performed as part of the act of publishing in a journal article or writing a thesis and can be very time consuming as it involves going back to ‘old’ data and familiarising yourself with it and possibly having to transform it into a more modern / recent required format. Included in this activity is the preparation of an ‘experimental’ description, which involves a description of the experiment with some operational values or parameters, references and standard operating procedures (and deviations from these). This work is performed by the researcher and the NCS publishes around 60 papers a year containing an average of 2 structures each. The average time per structure for this activity is:

	Before I2S2	*After I2S2*
• Tidy data	8 minutes	to be determined
• Format data	4 minutes	

and per report is:

• Write experimental	20 minutes

There is an almost linear relationship between the number of structures being presented and the time taken to generate supplementary data, however an umbrella experimental section can often be used.

6) Validate and Reuse

The measurements and benefits arising from being able to validate and reuse data are generally all qualitative. Firstly, in order to be able to reuse data one needs confidence in its correctness and therefore a significant amount of time can be saved if the re-user does not have to check the data before making use of it. Secondly, data that are made available in a published common format may be readily, or automatically, understood and reused or incorporated into other collections, again saving time in interpretation. Thirdly, data made openly available in a well-structured form are easy to find, saving time or opening up new opportunities. Finally, well-managed data (that are openly available) can persist for a longer time. This easy to find, persistent availability will result in more citations for a longer period of time.

3. I2S2 Benefits Use Case 2: a Researcher's Perspective

This use case utilises the I2S2 lifecycle activity model as a starting point for analysing the changes and benefits arising from I2S2 and will feed into a second stage I2S2 benefits case study. It is written from a researcher perspective (Martin Dove, University of Cambridge).

3.1 Value Chain and Impact Analysis

Activity	**Change in and Benefit from I2S2**	**Impact Weighting** (1 high- 5 low)	**From When?** (year1-5+ etc)	**Who Benefits?** (researcher, research group, institution, discipline, etc)
Research Concept	Access to previous data enables us to build on previous work much more easily.	2	Year 1	Researcher & research team; Discipline
Write Proposal	Access to previous data enables us to incorporate it into subsequent proposals.	4	Year 1	Researcher & research team
Peer Review Proposal	Probably low impact because reviewers don't typically have time to double check much, and then they tend to work from prior publications rather than data.	5	N/A	N/A
Start Project	Availability of previous work will help new projects have some reference point.	3	Year 1	Researcher & research team; Discipline
Acquire Sample & Collate Information	Information supplied by user in form that can be propagated through the system.	3	Year 1	Research team
Conduct Experiment	Sample information, such as dimensions and mass, and all run numbers, can be assigned to an experiment and available to all members of the research team. Reduces time latency in subsequent analysis.	1	Year 1	Research team

Activity	Change in and Benefit from I2S2	Impact Weighting (1 high- 5 low)	From When? (year1-5+ etc)	Who Benefits? (researcher, research group, institution, discipline, etc)
Process Raw Data	Full sample metadata record available. Available for preservation & curation and researchers to revisit and re-use quickly.	1	Year 1	Research team; ISIS or other facility
Analyse Derived Data	Full sample metadata record available. Available for preservation & curation and researchers to revisit and re-use quickly.	1	Year 2	Research team; ISIS or other facility
Interpret Results Data	Full sample metadata record available. Available for preservation & curation and researchers to revisit.	1	Year 2	Research team; Other researchers; ISIS or other facility
Documentation, Metadata and Storage	Majority of metadata already captured and automatically assigned.	1	Year 2	Research team; Other researchers; ISIS or other facility
Appraisal and Quality Control	I am guessing that in the overall aims this one might be a bit too ambitious to have a higher impact.	4	N/A	N/A
Programs (generate customized software)	Availability of wide suite of data for testing. In turn this is of significant benefit to the wider community, because the software is more robust and more widely applicable.	1	Year 2	Code developers; Research team; Discipline; ISIS or other Facility
Archive, Preservation and Curation	Committing to seamless and immediate archive. Benefit is that data are available for re-use by researchers, and for use in training of new users and by code developers.	1	Year 2	Researcher & research team; Other researchers; New users; Code developers; ISIS or other facility; Discipline

Activity	Change in and Benefit from I2S2	Impact Weighting (1 high- 5 low)	From When? (year1-5+ etc)	Who Benefits? (researcher, research group, institution, discipline, etc)
IPR, Embargo and Access	Management of embargo automatic.	3	Year 2	Researcher & research team
Write Usage Reports	Full data immediately available, leading to improved reports.	1	Year 1	Researcher; ISIS Facility
Submit to Publications Database	I think that this already appears to happen automatically within STFC.	N/A	N/A	N/A
Prepare Supplementary Data	Full data record available in structured form (automatic preparation?). Benefit of time saving for the author, and higher likelihood of comprehensive reporting.	1	Year 1	Research team; Discipline
Prepare Manuscript	Supporting data readily available. Benefit of time saving for the author.	1	Year 1	Research team
Peer Review Research	Full data record & provenance available for checking and validation. This will become increasingly important, but we are still at the “becoming” so without an obvious consensus emerging as to how to do this.	3	Year 3	Research team; Other researchers; Discipline
Publish Research	Automatic; Increased visibility and/or citation of data.	1	Year 1	Research team

Activity	Change in and Benefit from I2S2	Impact Weighting (1 high- 5 low)	From When? (year1-5+ etc)	Who Benefits? (researcher, research group, institution, discipline, etc)
Validate & Re-use (a) by original team (b) by others	Full data record & provenance available for re-use in new experiments; Full data and provenance available for preparing lectures, talks, and review papers; Full data and provenance available for generating new tools and developing existing programs; Full data and provenance available for training of other users, including making available extensive suites of data that users can select as best match their own interests; Published datasets are complete and accompanied by full metadata / documentation, which allows rapid assessment of validity, correctness and reusability.	1 + 1	Year 2	Research team; Code developers; New users; Other researchers; Training courses; Discipline

Key Resources?
Things we need to quantify to measure scale and impact of benefits and any comments?:

5. **Time spent on activity/facility before and after I2S2 changes**: Having good access to previous data should lead to significant savings on time, primarily by reducing time latency between activities. More than that, often times when the time cost in finding older data stops us re-using older data.
6. **Staff Salaries** (Principal Investigator, Administrator, Post-Doctoral Research Assistant, PhD Student average rates salaries incl. NI and pension costs): Within the university this sort of variable has only a loose connection with what people actually do. Thus universities have scope to think about cost savings in this regard: it is really only useful to think of increasing research effectiveness of staff. And in that sense the increased research effectiveness enabled by the significantly enhanced access to data and metadata outlined above will have an impact through better research and greater productivity measured in research outputs will lead to financial benefits through mechanisms such as grant proposal success and institution ranking.
7. **FEC on salaries** (assume average of 100%? Check re Administrator): Same comment as above.
8. **RAL Facility Cost** (Cost for Diamond and ISIS per day/hour and what this includes): Same comment as above. But as better research and greater productivity will in the long term lead to financial benefits for the institutes, the same will also be true for facilities.
9. **Size and % of Users/Community who will benefit from I2S2 changes**: The size of the user community is currently growing and we expect it to expand considerably in coming years (particularly if we succeed in our proposal to build a total scattering instrument at Diamond). In terms of beneficiaries from I2S2 I am pretty sure the percentage of users who will benefit will be close to 100%.

Measurable Impacts?
Any important impacts from I2S2 we can quantify and suggest metrics for:

1. **Time savings for researchers and support staff**: The way this is significant is that it removes the need to find older data from researchers and support staff. The result will be that some things that don't get done (i.e. some datasets wanted) will become automatic. In part there are time savings on colleagues who need to find data (maybe of order of an hour at a time), but the biggest impact is in terms of reducing the time latency from around one day currently to around five minutes. Latency can be a big hindrance to effectiveness, particularly as it impacts negatively on the researcher's workflow. In the worst case, the existence of latency in obtaining data may mean that the researcher is not able to return to the problem for several days when the time assigned for this work cannot be used effectively. In the bigger picture, I2S2 should mean that the time taken to reach publication is reduced considerably.

2. **Increased output** i.e. number of samples and/or datasets processed: the key thing here is the number of datasets that are analysed. Now there is always one cost that I2S2 can't impact, namely the time it takes to run the detailed number crunching (which takes several days), but where I2S2 can have an impact is in enabling different stages of the data analysis to be available in a useable form for researchers and their colleagues to easily pick up at any time. I think that the number of completed studies before and after I2S2 is a good measure. This can be measured by the number of papers that cite our RMC code, which is currently 15 per year.
3. **Improved application/efficiency of tool/program**: Without doubt the development of our tools is not helped by the lack of easily accessible data, particularly the scope of data. In particular, as we add new functionality – four immediate developments are for multiphase samples, inclusion of inversion molecular potential energy functions, increased ease of use for magnetic studies, and increased easy of use for x-ray studies – we need data from a number of different materials in order that the final tools are both robust and general. By analogy, codes such as the lattice simulation code GULP and the molecular dynamics code DL_POLY provide a wide range of test examples that are used for both developer and new-time user.
4. **Size of User Community for Facility and Tool/Program**: The total number of annual users at ISIS is fixed and can't be changed, BUT, the range of users can, and part of the benefit/impact will be in terms of the growth in the number of different users and the number who come back again. Moreover, we hope to see an expansion in the number of users from different facilities, particularly x-ray facilities, and in the breadth/depth of what they use the tools for. There are 263 people who have downloaded our code, although we haven't chased usage. This may be primarily for neutron work, and we want to see a significant expansion for people who use x-ray scattering. We anticipate the potential for very substantial increases in the number of users in this case, both with new instrumentation at facilities such as Diamond, and also the availability of lab-based diffractometers for this work (see Qualitative Impacts section below).
5. **Effectiveness of training**: The scope for training of new people will benefit by having a wide range of available data. We have worked quite hard on some of the training stuff, but more data is always needed. I had an idea once of making all our unanalysed but tame (no hidden nasties) data available for training purposes, so that new users not only learn how to use the tools but also get out new research work at the same time. I have not heard of user groups using data this way, and it could be an interesting thing to try. One metric in this case would be to question trainees via an evaluation form at the end of training workshops as to whether they could see the value of the tools through the data to which they were exposed.
6. **% of research data from experiment made available for re-use**: I would like to see us aim at 100% for this. This is to be compared with a current number of close to 0%. Some of the impact will be anecdotal. For example, I have colleagues who want to use our tools to analyse x-ray data. We have done work with x-ray data before, but I don't have at hand prior data, so need to ask my colleague in Oxford for examples and advice, but with the data available we would be able to work independently.

To conclude, the one thing I would like to see at the end of this project for my use case, against which impact can be measured, is a system in place that enables us to store data from all stages of data analysis in a form that is properly usable. The test of usability is that someone new can use stored data for training with only a minimum of help.

Qualitative Impacts?

Other important impacts we believe I2S2 will deliver but will be difficult to measure due to timescales or other factors:

1. **Industry**: There is a company that now markets x-ray PDF systems and who could use data of the sort in this case study. Bringing them on board and enabling them to utilise the system for x-rays would be good. We are in discussion about forming a formal collaboration, with the aim that our tools are used on the companies equipment and bundled with their software.
2. **New Facilities**: There is a strong possibility of a new PDF machine at Diamond comparable to facilities currently only available in the USA, which would create completely new opportunities for UK researchers; the outcome of the proposal will be known in March. Having I2S2 systems in place from the outset should this be funded would be a good impact.
3. **I2S2 Persistent Citation**: Having some means by which data have some sort of "perpetual" URI for citation in papers would enable other workers to make use of data that feature in scientific publications. This is something that is now being taken forward in the SageCite project.
4. **New Training Courses:** Matt Tucker (one of the code developers and working at ISIS) runs one or two training courses a year, and we are discussing holding a longer training course as a regular point in the national diary for crystallography research. Having an extensive set of managed data for this would be great. I have a secret ambition to give our unanalysed data for training purposes, with the idea that training could merge into managed research with publication outputs. Note that other tools that have data training suites. A good on-line example is from http://www.mrc-lmb.cam.ac.uk/harry/imosflm/ver104/documentation/tutorial.html; the aforementioned GULP and DL_POLY both ship with suites of test files.
5. **Future Extensions**: When we have a system in place, I would like to aim at 5 new data sets per year from the core team to be uploaded (one data set will include several different runs at various temperatures and conditions), with new collaborators contributing the same rate. I also want to see us developing a new x-ray test suite.

Impact Timescales: Time limitations on the project mean that many of these will not be measured over the time scale of the project, in part because of latency (e.g. time to publication). The right time for the impact analysis is probably in about five years' time, allowing for some of the benefits to be better embedded within the community and the scale of the impact to be visible.

Maximum Selection of Five Activities by Impact for Further Analysis

I have flagged more than five #1 areas above. To reduce to five I would like to merge things a bit, recognising that in so doing I need to be careful not to merely increase the number of activities by implication.

1. Enabling data re-use by the researchers and others
2. Capturing information at the time the experiment is conducted
3. Storing processing and analysing of data at each stage of the process
4. Having all information available for writing papers and supplementary information document.

5. Having data available for code development and for training

I recognise that there are broad overlaps between some of these, for example 1 & 5 could be the same thing, and 2 & 3 are needed for 4 as well as needed for 1 & 5.
