Model of the data continuum in Photon and Neutron Facilities

PaN-data ODI

Deliverable D6.1

Grant Agreement Number	RI-283556
Project Title	PaN-data Open Data Infrastructure
Title of Deliverable	Model of the data continuum in Photon and Neutron Facilities
Deliverable Number	D6.1
Lead Beneficiary	STFC
Deliverable Dissemination Level	Public
Deliverable Nature	Report
Contractual Delivery Date	30 Sept 2012 (Month 12)
Actual Delivery Date	

The PaN-data ODI project is partly funded by the European Commission under the 7th Framework Programme, Information Society Technologies, Research Infrastructures.

Abstract

This report considers the potential for data management beyond the management of raw data to record, link, combine and publish information about other data, digital objects, actors and processes involved in the whole facilities science lifecycle – broadly covered by the term *provenance* of information.

In particular, the report will consider:

1. The *data continuum* involved in the lifecycle of facilities science, considering the stages undertaken in the lifecycle, the actors and computing systems typically involved at each stage, and the metadata required to capture the information at each step.
2. Consider a specific but representative example of a scientific lifecycle within facilities science and discuss its consequences for practical data management including provenance in facilities
3. Consider a number of other specific examples where parts of the scientific lifecycle can be given additional support to derive additional benefit for facilities infrastructure staff and facilities users.

Keyword list

Data analysis, data continuum, provenance, research lifecycle, research output, workflow

Document approval

Approved for submission to EC by all partners on 12.11.2012

Revision history

Issue	Author(s)	Date	Description
0.1	Brian Matthews (STFC)	04 Sept 2012	First Draft
0.2	Brian Matthews (STFC), George Kourousias (ELETTRA), Erica Yang (STFC)	26 Oct 2012	Complete draft including scenario descriptions
0.3	Brian Matthews (STFC)	31 Oct 2012	Reworked section 2.
0.4	Brian Matthews (STFC)	1 Nov 2012	Added conclusions section, references
0.5	Brian Matthews (STFC), Tom Griffin (ISIS)	9 Nov 2012	Revised and additional scenario descriptions. Comments from Frank Schluenzen (DESY) and Catherine Jones (STFC)
1.0	Brian Matthews (STFC)	12 Nov 2012	Final version

Table of contents

Page

Executive Summary

When considering how to provide infrastructure to support facilities-based science, it is helpful to consider the whole of the research lifecycle involved, from submitting applications for use of the facility, through sample preparation and instrument configuration and calibration, through data acquisition and storage, secondary data filtering, analysis and visualisation to reporting within the research community, informally and through formal publication. By taking an integrated approach, taking into account the provenance of the data (Creation, Ownership, History), the infrastructure can maximise the potential for science arising from the data.

In general, there is a *Data Continuum* from proposal to publication where data and metadata can be managed together as a record of the experimental lifecycle of an experiment. This lifecycle goes through the stages as follows.

1. **Proposal:** The user submits a proposal applying to use a particular instrument on the facility for time to undertake experiments on particular material samples. This is lodged with the Facility.
2. **Approval:** the application is judged on its scientific merits and technical feasibility of the proposal, successful proposals being allocated a time period within an operating cycle of the instrument.
3. **Scheduling:** Time on the instrument is allocated to successful proposals to determine when the experiment will scheduled to take place.
4. **Experiment:** During a visit to the facility, a set of samples are placed in the beam and a series of measurements are taken. Different instruments at the facilities have their own characteristics, but all have data acquisition software which will take data on the parameters of interest.
5. **Data Storage:** Data is aggregated into data sets associated with each experiment, stored in secure storage, within managed data stores in facility, and systematically cataloged.
6. **Data Analysis:** The scientist takes the results of the experiments (the “raw data”), and carries out further analysis. The data from the instruments is typically in terms of counts of particles at particular frequencies or angles, and needs highly specialized interpretation to derive the required end result, typically a “picture” of a molecular structure, or a 3-D image of a nanostructure.
7. **Publication:** a suitable scientific result having been derived from the data collected, then the scientist will report the results within journal articles. The facility would usually like to be acknowledged and informed of its publication, so that it can track the impact of the science derived from the use of its facilities

Early stages in the process are relatively speaking within the facility’s control and using the facility’s staff and information systems and thus it is relatively straightforward to provide integrated support for those stages of the process. Later stages (analysis and publication) are largely outside the control of the facility, and thus are hard to contain within a single provenance management system. This leads to a careful consideration of the value and costs of managing this information.

Provenance is still an experimental area within PaN-data, with not all partners regarding it as a core part of the infrastructure, but rather within the scientific user community, and not necessarily delivering benefits which outweigh the additional costs in storage, tooling and expertise, as shown

in the user survey [PaN-data-Europe D7.1]. Providing a universal solution to provenance is a difficult problem, and is probably too complex and expensive at this stage.

Nevertheless, provenance information is potentially of great value, and in scenarios where provenance can be captured and utilized effectively within the facilities data management infrastructure, and with identifiable additional cost, it can make the scientific process more efficient and lead to better science. Thus the use of provenance is scenario dependent; in this work package, we are identifying scenarios where we can apply provenance techniques and demonstrate additional value from its use.

The initial scenarios considered are:

- The TwinMic X-ray spectro-microscope beamline at Elettra, This case study is considering the complex interactions between different stages of experiment preparation, execution and post-processing which are involved in a multi-visit experiment (e.g. one which takes place over more than one allocation of experimental time), which requires a higher level of coordination and support.

- The SANS2d Small angle neutron scattering instrument at ISIS, which seeks to automate the “near to experiment” processes in the experimental cycle, which involve experiment setup and execution, post-processing to provide “reduced” data, which is a fairly routine data analysis step, and publication of results via an electronic notebook.

- X-Ray tomography experiments at the Diamond Light Source, which have particular intensive data handling requirements to process the images captured from the beamline instruments, into a reconstructed 3D model. The sheer size and number of such reconstructions mean that there are special issues of data handing and processing which are best handled within a systematic data management infrastructure.

- GEM Xpress (“measurement-by-courier”) service for powder defraction at ISIS. This scenario is an example of a mode of use of a facilities instrument where the involvement of the experimental team is at a minimum. The experimental team does not visit the facility but sends the samples and the experiment is carried out by the instrument scientist and reduced data returned to the experimenters. Thus whole process remains in the facilities control and amenable to tracking and automation.

- Using publication and data catalogues within the ISIS infrastructure to track research outputs, including publications and final resultant data. This would provide an enhanced service for users to increase output availability, and allow the facility to more accurately assess research impact.

These scenarios show that there are clear cases (and there are further ones which could also be explored) where tracing provenance is of value, and thus generic tools, if they can be developed within reasonable cost, could be explored within PaN-data, which can be used to support such scenarios.

1 Introduction

1.1 Background: facilities science

Neutron and photon sources are a class of major scientific facilities serving an expanding user community of 25,000 to 30,000 scientists across Europe, and a much wider community across the world, within disciplines such as crystallography, materials science, proteomics, biology and even archaeology

The traditional approach of many of the facilities leaves data management almost entirely to the individual instrument scientists and research teams. While this local responsibility is well handled in most cases, this approach in general has become unsustainable to guarantee the longevity and availability of precious and costly experimental data. Large-scale facilities are advanced scientific environments which have demanding computing requirements. Modern instruments can generate data in extremely large volumes, and as many instruments as possible are placed around target areas or beam-lines in order to maximize the output from the expensive neutron or synchrotron x-ray resource. Consequently, the data volumes are large and increasing, especially from synchrotron sources, and the data throughput is very high, and thus the data management requires large-scale data transfer and storage. The diverse communities involved in building instruments and software and also the different academic communities and disciplines, has lead to a proliferation in data formats and software interfaces. This increased capability of modern electronic detectors and high-throughput automated experiments, means that these facilities will soon produce a "data avalanche" which makes it essential that a framework be developed for efficient and sustainable data management and analysis.

Not only is this becoming unfeasible considering the dramatic increase in size of some of the data sets, it is also counterproductive as a way of managing the workflow of the science through the facility. Today's scientific research is conducted not just by single experiments but rather by sequences of related experiments or projects linked by a common theme that lead to a greater understanding of the structure, properties and behaviour of the physical world. These experiments are of growing complexity, they are increasingly done by international research groups and many of them will be done in more than one laboratory. This is particularly true of research carried out on large-scale facilities such as neutron and photon sources where there is a growing need for a comprehensive data infrastructure across these facilities to enhance the productivity of their science.

The data collected has a large number of parameters, measured both from the operating environment (e.g. temperature, pressure) and from the sample (typically angles from a scattering pattern) and this requires a multi-variate analysis, typically over several steps. To handle the data volumes and to use bespoke software, distributed computation such as Grid or cloud systems are required to access high-performance computation.

Facility users are typically from university research groups, but also from a number of commercial organizations such as pharmaceutical companies, and in both cases the data can be sensitive. Consequently, there is a need to manage different data access requirements, sharing data with a research team in different institutions, and restricting access to non-authorised individuals.

Finally, as expensive investments (e.g. DLS cost some £400M to commission), governments wish to maximise the science output from facilities. Thus there is a need to maximise the use of data for the original data collectors, by capturing, organising and presenting it to them in a manner so that it can be analysed with the most up-to-date techniques, and not be a subject of unnecessarily repetition of the experiment through lost or poor quality data. Further, there is an increased recognition that output can be maximised by managing data for the long-term so that it can be reused by future scientists rather than re-doing the experiment.

Thus when considering how to provide infrastructure to support facilities-based science, it is helpful to consider the whole of the research lifecycle involved, from submitting applications for use of the facility, through sample preparation and instrument configuration and calibration, through data acquisition and storage, secondary data filtering, analysis and visualisation to reporting within the research community, informally and through formal publication. By taking an integrated approach, taking into account the provenance of the data (e.g. Creation, Ownership, History), the infrastructure can maximise the potential for science arising from the data.

Consequently, the facilities have a strong requirement for a systematic approach to the management of data across the lifecycle.

1.2 Scope of this report

The management of data resulting from the experiment is considered and handled via data catalogues in PaN-data ODI WP4. This report considers the potential for data management beyond the management of raw data to record, link, combine and publish information about other data, digital objects, actors and processes involved in the whole facilities science lifecycle – broadly covered by the term *provenance* of information.

In particular, the report will consider:

1. The *data continuum* involved in the lifecycle of facilities science, considering the stages undertaken in the lifecycle, the actors and computing systems typically involved at each stage, and the metadata required to capture the information at each step.
2. Consider a specific but representative example of a scientific lifecycle within facilities science and discuss its consequences for practical data management including provenance in facilities.
3. Consider a number of other specific examples where parts of the scientific lifecycle can be given additional support to derive additional benefit for facilities infrastructure staff and facilities users.

We will not in this report consider: access control, except when noting that specific actors are involved in the stages of the process; technical standards; description of proposed general architecture, models or ontologies; or specific tools for managing provenance, workflow or data management. Some of that material is covered in other work packages or subsequent deliverables of this work package.

2 Data Continuum for Facilities

2.1 Overview of facilities lifecycle

We consider a simplified and idealized view of the stages of the science lifecycle within a single facility, as illustrated in
Figure 1.

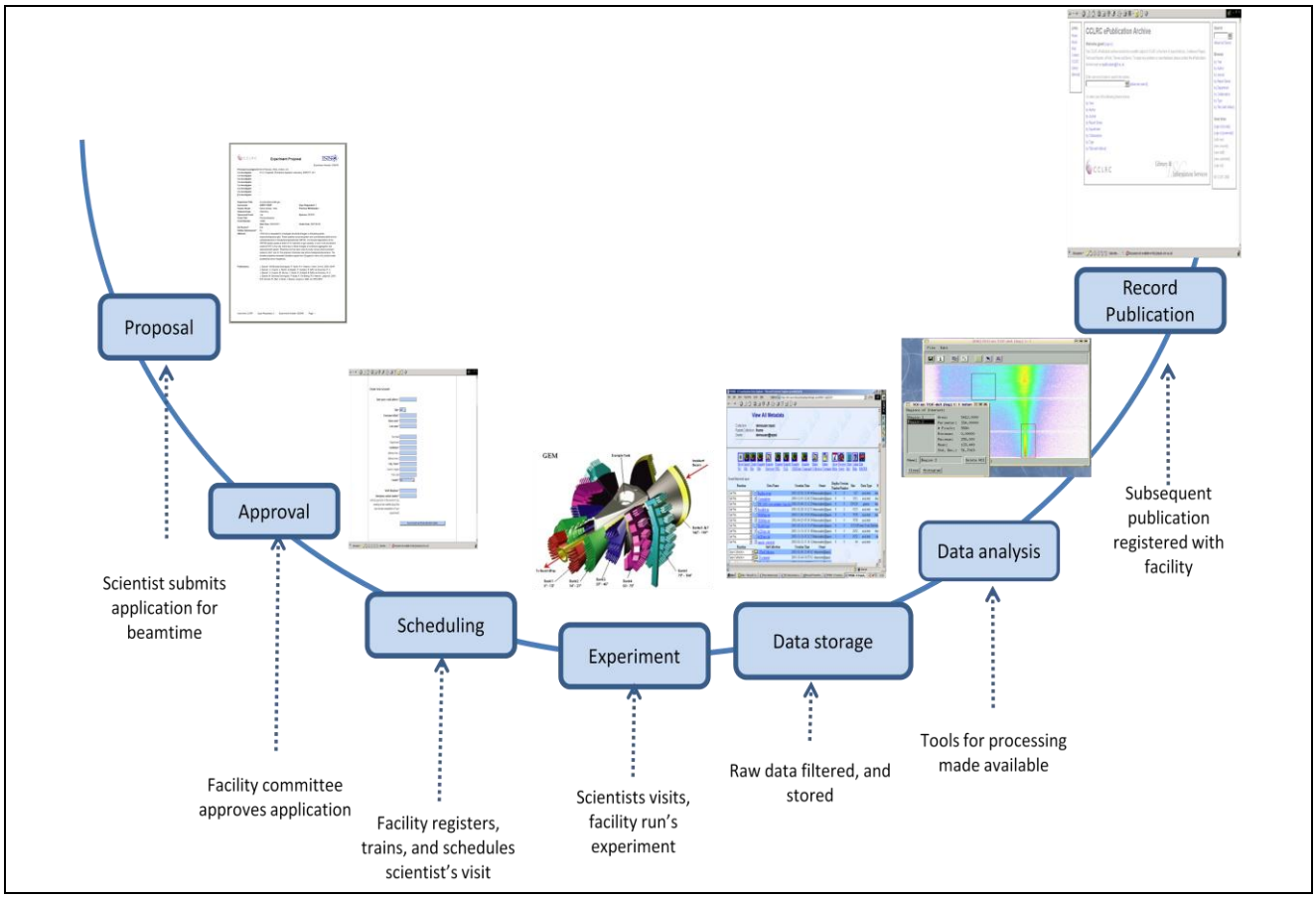


Figure 1: an idealised facilities lifecycle

Thus in general, these stages are as follows.

1. **Proposal:** The user submits a proposal applying to use a particular type of instrument on the facility for time to undertake experiments on particular material samples. This is lodged with the Facility.
2. **Approval:** the application is judged on its scientific merits and technical feasibility of the proposal, successful proposals being allocated a time period within an operating cycle of the instrument.
3. **Scheduling:** Time on the instrument is allocated to successful proposals to determine when the experiment will scheduled to take place.
4. **Experiment:** During a visit to the facility, a set of samples are placed in the beam and a series of measurements are taken. Different instruments at the facilities have their own characteristics, but all have data acquisition software which will take data on the parameters of interest.

5. **Data Storage:** Data is aggregated into data sets associated with each experiment, stored in secure storage, within managed data stores in the facility, and systematically cataloged.
6. **Data Analysis:** The scientist takes the results of the experiments (the "raw data"), and carries out further analysis. The data from the instruments is typically in terms of counts of particles at particular frequencies or angles, and needs highly specialized interpretation to derive the required end result, typically a "picture" of a molecular structure, or a 3-D image of a nanostructure.
7. **Publication:** a suitable scientific result having been derived from the data collected, then the scientist will report the results within journal articles. The facility would like to be acknowledged, citing the instrument used, and informed of its publication, so that it can track the impact of the science derived from the use of its facilities

Thus there is a *Data Continuum* from proposal to publication where data and metadata are managed together as a record of the experimental lifecycle of an experiment. .

2.2 Actors involved in the lifecycle

Different people are involved at the various stage of the lifecycle. The major actors involved in the lifecycle include:

- **The Experimental Team**: a group of largely external (e.g. University) researchers who propose and undertake the experiment. This team would typically be led by a **Principal Investigator** and would have expertise on the sample under examination within the experiment, its chemistry and properties. They may have some knowledge of the analytic technique being used to perform the experiment (e.g. crystallography, small-angle scattering, powder diffraction), but typically would not have detailed knowledge of the characteristics of the instrument, relying for this on assistance from the instrument scientist.

- **The User Office**: a unit within the facility dedicated to managing external users of the facility. User Office staff and systems will typically register users, process their applications for beam-time, guide them through the process of visiting and using the facility, including managing any induction or health and safety processes, and collate information on the scientific outputs of the visit.

- **The Instrument Scientist**: a member of the facility's staff with specialist scientific knowledge of the capabilities of a particular instrument or beam-line and its use for sample analysis. The will typically advise and assist with the experiment on the instrument and often are included within the experimental team.

Other actors involved may include:

- **Approval panels**, formed by scientific peers and charged with assessing proposals and allocating time on the instruments;

- **Facility libraries**, which may collect information on resulting publications;

- **Facility infrastructure providers:** who maintain computing and data infrastructure within the facilities; and

- **Facility operations staff**: who manage the physical operation of the facilities, the moving of equipment, handling samples and chemicals, running the facility's source of beam.

Note that from the perspective of PaN-data, we can distinguish between *internal users* of the computing and data infrastructure, including the user office managers and instrument scientists on the facilities staff, and *external users,* which are the end-users of the facilities, which typically come from universities and other research institutions. Both are users of the computing and data infrastructures, the internal users using the infrastructures on a day-to-day basis, and the external users who interact with the infrastructure to expedite their work through the system and generating the results. Thus both of these groups have a stake in the infrastructure and PaN-data thus maintains strong links with both groups:

- **Internal users**: facility staff who are within the same organisation and have daily interactions with user office and instrument scientists.

- **External users**: facilities maintain very close working relationships with their user communities, through their normal operations, often working with the same experimental teams. Further, facilities have frequent consultative activities with external users, such as user group meetings, newsletters, mailing lists etc. Consequently, facilities have close knowledge of the needs and priorities of external users.

2.3 Stages of the experimental lifecycle in detail

These stages are considered in detail below. In each stage, we give an indication of:

- **Actors**: The people involved in each stage of the process, and their role in that stage

- **Sub-processes**: an idealized breakdown of the stage into some general sub-stages of the processes and their interactions and dependencies. We give a schematic workflow diagram of these stages. Note that some sub-stages are undertaken without the necessary participation of the facilities staff; these are part of the users' scientific workflow rather than that of the facility. These are signified in the diagrams by dashed lines and boxes.

- **Information Systems:** The computer systems which typically are involved in supporting data and metadata management at each stage of the process.

- **Data:** The scientific data involved at each stage.

- **Metadata:** The major categories of metadata which can be used to characterize the activities and data collected at each stage.

Note that this is an *idealized* description of the process undertaken within a facility; there are likely to be many exceptional cases and deviations, or cycles, stages undertaken in different order. Indeed, any particular instance of an experiment may well deviate in some aspect to this idealized view. Nevertheless, we feel that it useful and instructive to develop this idealized view so that we

can identify the general information systems and data and metadata sources which we can use within an integrated and federated data infrastructure.

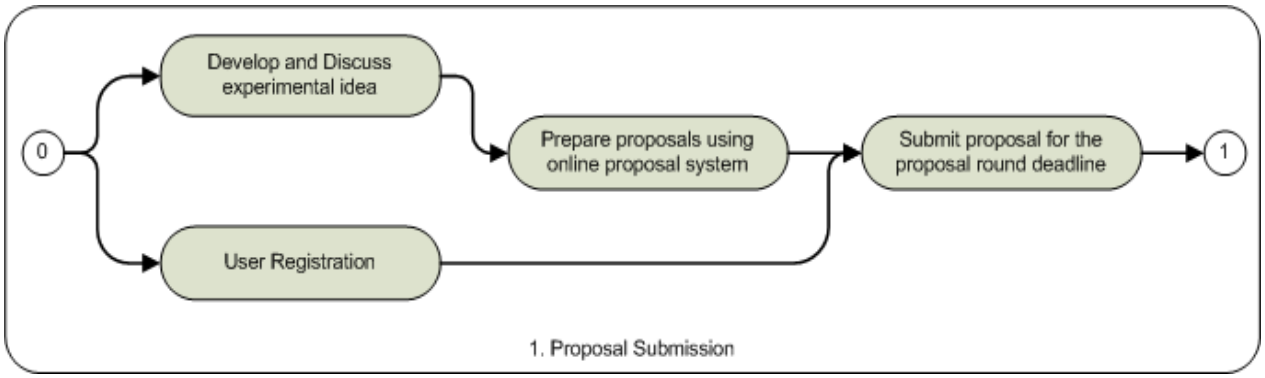
2.3.1 Proposal

Description.

The user submits a proposal applying for beam-time, to use a particular instrument on the facility for a period of time to undertake a number of experiments on particular material samples under particular conditions. This proposal outlines the intention of the experiment, with an assessment of the likely value of the results and a description of the prior expertise of the experimental team. Practical information concerning the safety and justification of the choice of instrument will also be included. This will be lodged with the Facilities User Office, who will register new users and maintain their record.

Sub-processes

A proto-typical proposal submission process would be as follows[1].



The proposal stage would have the following sub-stages:

- **Formulating a proposal idea:** this is the development of the idea for an experiment at a facility. Users are encouraged to discuss this with the instrument scientist staff at the facility to identify the most appropriate instrument and technique to maximise the chances of getting the best scientific result.
- **User registration:** The proposal submitters will need to register with the user office to gain access to the submission system (typically this will only need to be on the first submission).
- **Proposal preparation:** proposal is prepared by principal investigators via the online submission system. Again guidance from the facilities staff may be sought.
- **Proposal submission:** Proposal submitted via the online submission system before the round deadline

[1] See for example the advice on the ISIS website: http://www.isis.stfc.ac.uk/apply-for-beamtime/writing-a-beam-time-proposal-for-isis4408.html

Actors

- Principal Investigator : prepares and submits the proposal
- Instrument scientists : consults on the most appropriate experimental scenario
- User Office: registers users, ensuring their uniqueness; receives and processes the proposal

Information Systems

- User office systems,
- User registration and management,
- User identity,
- Proposal systems

Metadata Types

- user identity,
- instrument requested
- funding sources (e.g. research grant, funding councils, commercial contract etc).
- user institution (e.g. the institution the user is affiliated to
- sample description (e.g, description of chemical and its state).
- proposed experimental conditions (e.g. parameters temperature, pressure, measuring time)
- safety information. (e.g. explosive, radio-active, bio-active, or toxic substances; kept under extremes of temperature or pressure)
- experiment description, with a science case
- prior art (e.g. previous publications, preliminary investigations using laboratory equipment)

2.3.2 Approval

Description

The application goes to an approval committee who judges the scientific merits and technical feasibility of the proposal and makes a recommendation to approve or reject the proposal.

Sub-processes

The approval stage would have the following sub-stages:

- **Collating submissions:** The user office will collate the proposals which have been submitted in for a particular round (a deadline set for proposals for experiments for a particular period of facility operation[2].).
- **Proposal Evaluation:** The approval committee will be convened to consider and adjudicate on the submissions for the round. This may include recommending the use of alternative instruments.
- **Informing Results:** The results of the adjudications will be conveyed by the user office to the applicants.

Actors

- User Office: collates and convenes the approval panel; informs the results.
- Approval Panel: considers and adjudicates on the proposals

Information Systems

- User Office Systems,
- Proposal Systems

Metadata Types

- User identity,
- funding sources,
- experiment description
- proposals
- prior art

2.3.3 Scheduling

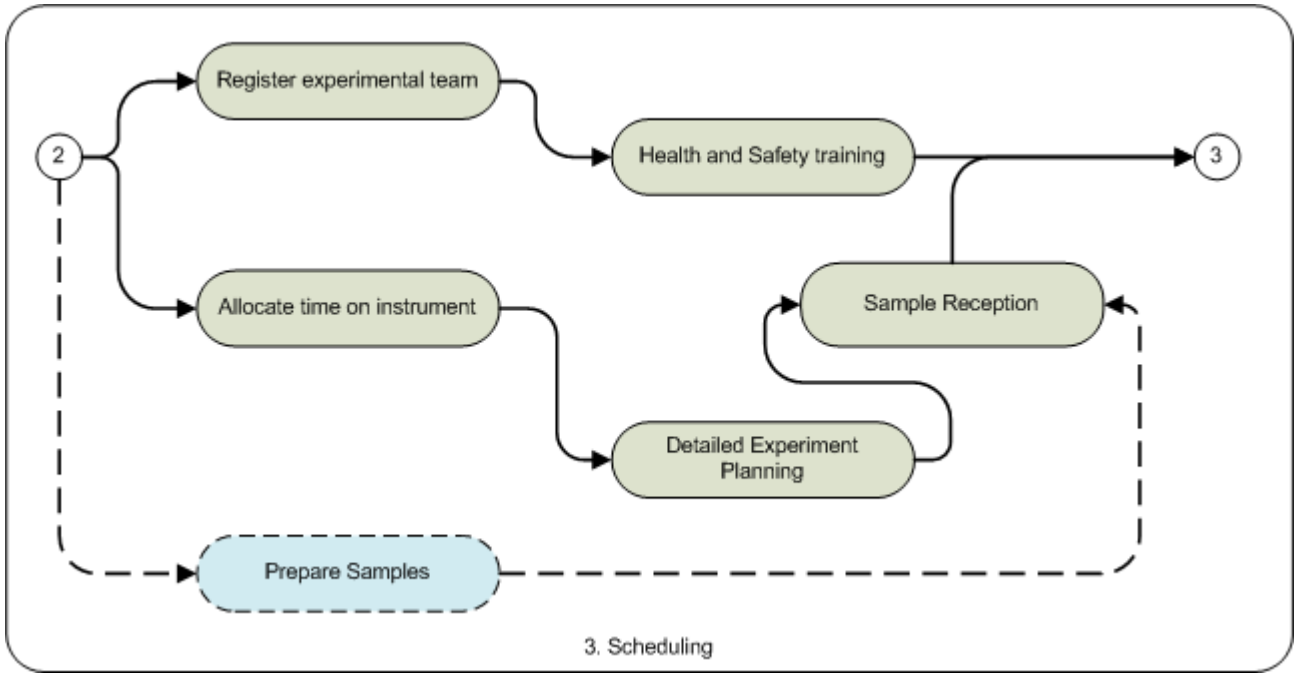
Description

Successful proposals are allocated a time period within an operating cycle of the instrument, and the experimental team prepare for their visit to the facilities site. At this time, there is a safety assessment of the proposed experiment: such experiments are frequently performed on dangerous materials (e.g. explosive, toxic, corrosive, radioactive, bio-active) and at extreme conditions (e.g. at

[2] Large-scale facilities have regular cycles of active operation and shut-downs, periods where no experiments are performed when maintenance and upgrades can be undertaken.

extremely high or low temperature, extremely high or low pressure). Therefore there has to be an evaluation of the correct handling of the material to ensure the safe procedure of the experiment. Further, there will typically be training of the experimental team on the safe and effective use of the hardware and software of the instrument.

Sub-processes



The scheduling stage would have the following sub-stages:

- **Allocate time on instrument:** the date and time and duration of the allocation of usage of an instrument will be scheduled. This may be a contiguous block of time, or a series of separate times at different dates.
- **Register experimental team:** those members of the team not already registered will need to be registered (e.g. research students and assistants, who may not be included on the proposal submission, but are expected to undertake the experiment as part of their research).
- **Training:** the experimental team will undergo training, especially in the safe use of the instruments. Facilities typically expect that this training will be carried in advance of the actual experimental visit to the facility (e.g. online or during a pre-visit).
- **Detailed experimental planning:** details of the samples and the experimental techniques to be undertaken will be planned by the team as much as is possible. Requirements for special handling of samples will be planned. Administrative issues, such as travel and accommodation will be covered.
- **Sample Preparation:** the experimental team will prepare the samples for analysis in the experiment, via chemical synthesis, crystallization, sample collection or other discipline de-

pendent methods. This is likely to be a major area of intellectual input of the experimental team (representing a major contribution to a doctoral thesis for example),and may take a great deal of time and intellectual effort, and expense, to prepare what may be a small and fragile sample. Thus this stage typically takes place in the university laboratory and the facilities teams have relatively little input[3] in the sample preparation process.

- **Sample Reception:** Samples frequently require special handling (e.g. maintaining low temperatures, high pressure, toxic or radio-active material), and are thus often delivered separately to the facility. This needs to be coordinated with the managers of operations at the facilities.

Actors

- User Office: register users, manage H&S training, schedule visit
- Experimental Team: prepare sample, plan experiment, undertake training
- Instrument scientist: plan experiment, schedule facilities access time,
- Facility operations: handling equipment and special requirements, handled samples.

Information Systems

- User Office Systems,
- H&S systems,
- Scheduling systems
- Sample tracking systems

Metadata Types

- User identity,
- Sample information,
- Instrument information,
- Experiment planning
- Safety information

2.3.4 Experiment

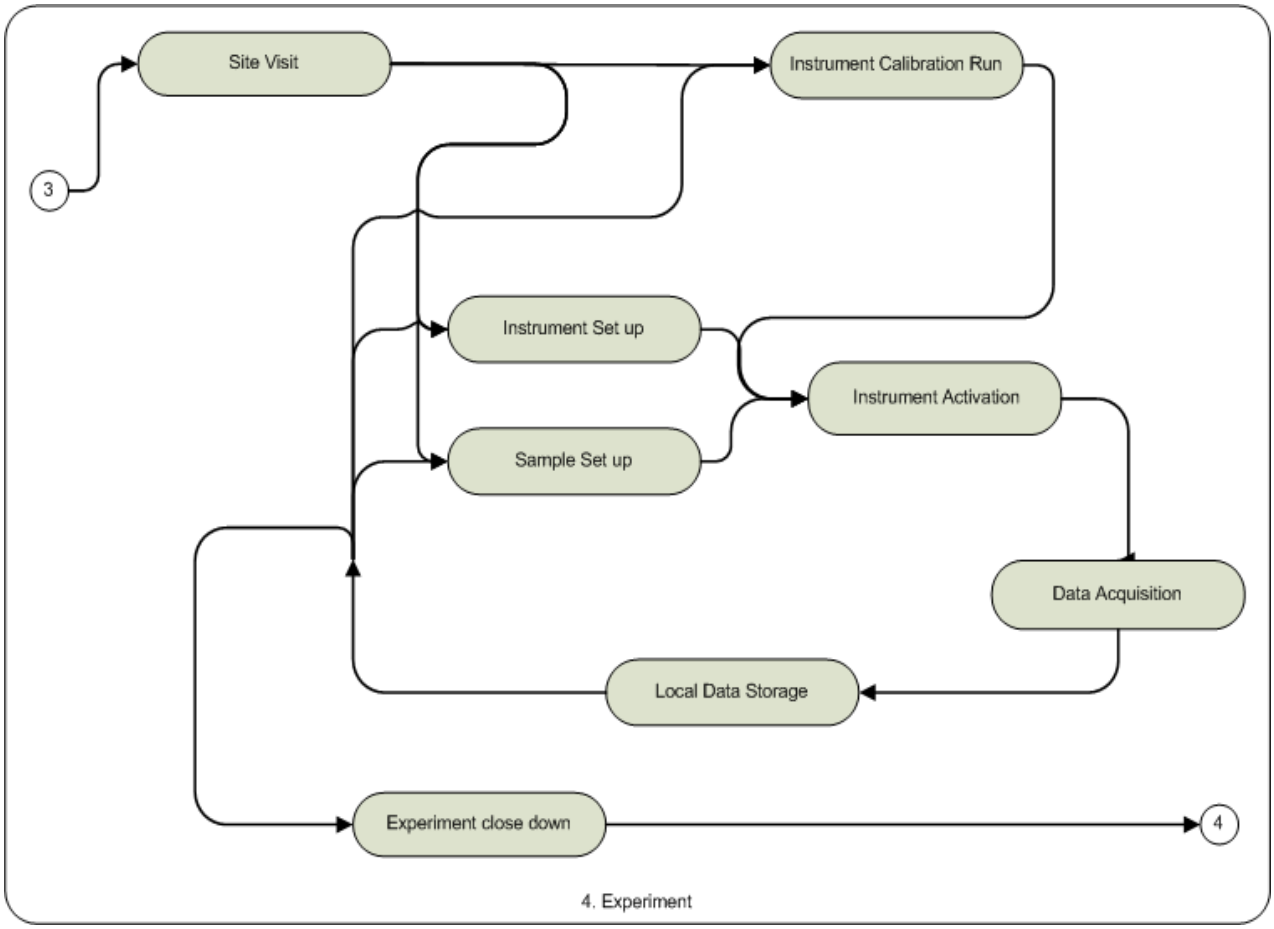
Description

During a visit to the facility, a sequence of samples is placed in the beam and a series of measurements are taken using the detectors. Different instruments at the facilities have different characteristics, but all will have data acquisition software which will take data measuring those parameters of interest measured by the instrument. This will be generally collected in a series of data files, named using some naming convention and in a format specific to the instruments, though

[3] At least in their facilities role; in practice, many facilities scientists have a role (and often joint appointments) as part of scientific teams in universities or other research laboratories; but in this report, we are considering them in their capacity as facilities support staff.

there is an effort to ensure that this is now collected in standard formats. Historically, this data is collected within the file systems associated with the instrument under the management of the instrument scientist. However, as data volumes have increased, there has been an increasing need to provide systematic support for this activity.

Sub-processes



This is a stage in the process which is difficult to generalize, as each experiment is likely to take a different course, there is likely to be much error and backtracking, changing parameters and conditions and samples, and rerunning the experiment. Nevertheless, we here try to capture the major steps undertaken in an idealized experiment.

The experiment stage would have the following sub-stages:

- **Site visit:** the experimental team visits the site and begins their experiments at their allocated time. This would require assembling the team, samples and any additional equipment required.
- **Instrument calibration:** typically an instrument calibration run, often against a reference sample will be undertaken. This could be taken at different intervals depending on the instrument (as little as once in a operating cycle, or repeatedly during a experiment). Instrument characteristics changes over time, parts become faulty, environmental conditions can affect the data collection, systematic errors can be included , so by taking reference data,

the results can be calibrated against a back ground result.

- **Instrument set up:** the environmental parameters, specialized equipment and measured characteristics can be adjusted for a particular run of the instrument. These may be changed repeated between measurements (e.g. to measure the same sample at different temperatures or pressures).
- **Sample set up:** a sample prepared into the final desired state, and needs to be mounted in the target area of the instrument.
- **Instrument activation:** when the sample and instrument are set up as desired, the beam is fired at the target sample for the desired length of time.
- **Data Acquisition:** during the instrument activation, data is streamed off the instruments;
- **Local data storage:** the data acquired is typically stored locally to the instrument, before being moved to a more permanent data store. In practice, there may be some initial data processing at this stage to see an initial view of the results, an evaluation of the data quality, potentially a visualisation to get a idea of how "good" the data which has been collected and potentially an opportunity to try again to collect better data.
- **Experiment close down:** the instruments are closed down, the samples cleared away (again with appropriate handling) specialist equipment removed.

With a number of samples being analysed within a period of allocated experimental time at different conditions and with retries when things go wrong, there are likely to be many cycles round these stages, so as emphasized this is a schematic view of this process.

Actors

- Experimental Team: Undertake the experiment
- Instrument scientist: assist the experimental team on undertaking the experiment.
- Facility operations: provide support for handling equipment and samples, and operating the facility.

Information Systems

- Sample tracking,
- Instrument control,
- Environmental monitoring,
- Data Acquisition systems,
- Data Management systems
- Electronic notebook systems

Data types

- Data sets of raw experimental data associated with each sample
- Calibration data

Metadata Types

- User identity,
- Sample information,
- Instrument information,
- Experiment planning,
- Environmental parameters
- Calibration information

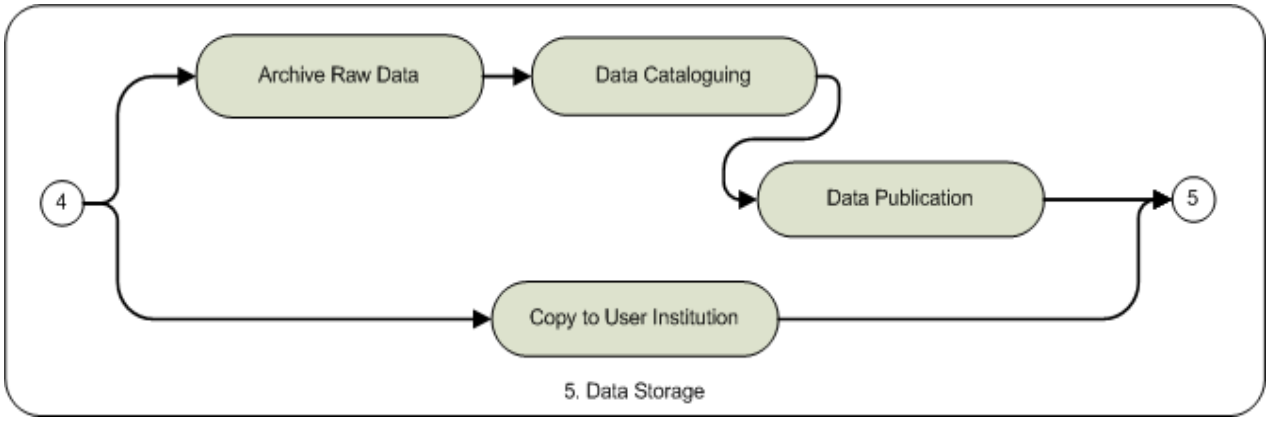
- Laboratory note books.

2.3.5 Data Storage

Description

Data is aggregated into data sets associated with each experiment and stored in secure storage, within managed data stores in facility and often for backup elsewhere. Additionally, with the increase in the systematic management of the data, this may be catalogued in a database. The data is kept there and made available to the user, typically for a period of time. There is increasing recognition that there is a need to retain this data potentially for a long period of time.

Sub-Processes



The data storage stage would have at least the following sub-stages:

- **Archiving the Raw Data:** data is moved off the data acquisition and storage local to the instrument onto a larger "live-data" online storage; possibly it will also be copied onto a archival system for long-term preservation of the data (kept separate from the live data).
- **Data Cataloguing:** A data catalogue entry of the data to be made, linking the raw data with parameter information from the experiment and to information on the user and context taken from the proposal.
- **Data publication:** Data is made remotely accessible. Access to data is subject to embargo, so data might not be *openly* accessible immediately. Assigning a persistent identifier to data and referencing the identifier in a publication would usually require immediate release of the data.
- **Copy to user institution:** data is optionally copied to the users' home institution; historically this has been done via tapes or disks to take data off site

In practice, it is likely that some of the stages in the data storage stage would be interleaved with the data acquisition and local storage; these processes may be done in real time while the experiment is being undertaken, depending on the amount of automation which has been set up. How-

ever, for convenience we separate them out.

Actors

- Experimental Team: arranging to take data off site.
- Data infrastructure team: managing the data storage and publication process.

Information Systems

- Data Acquisition systems,
- Data Management systems
- Data storage systems,
- Data publication systems
- Archival Systems

Metadata Types

- Data set information,
- File identifiers
- Instrument parameters,
- Preservation Description information,
- Representation Information.
- Persistent identifiers

2.3.6 Data Analysis

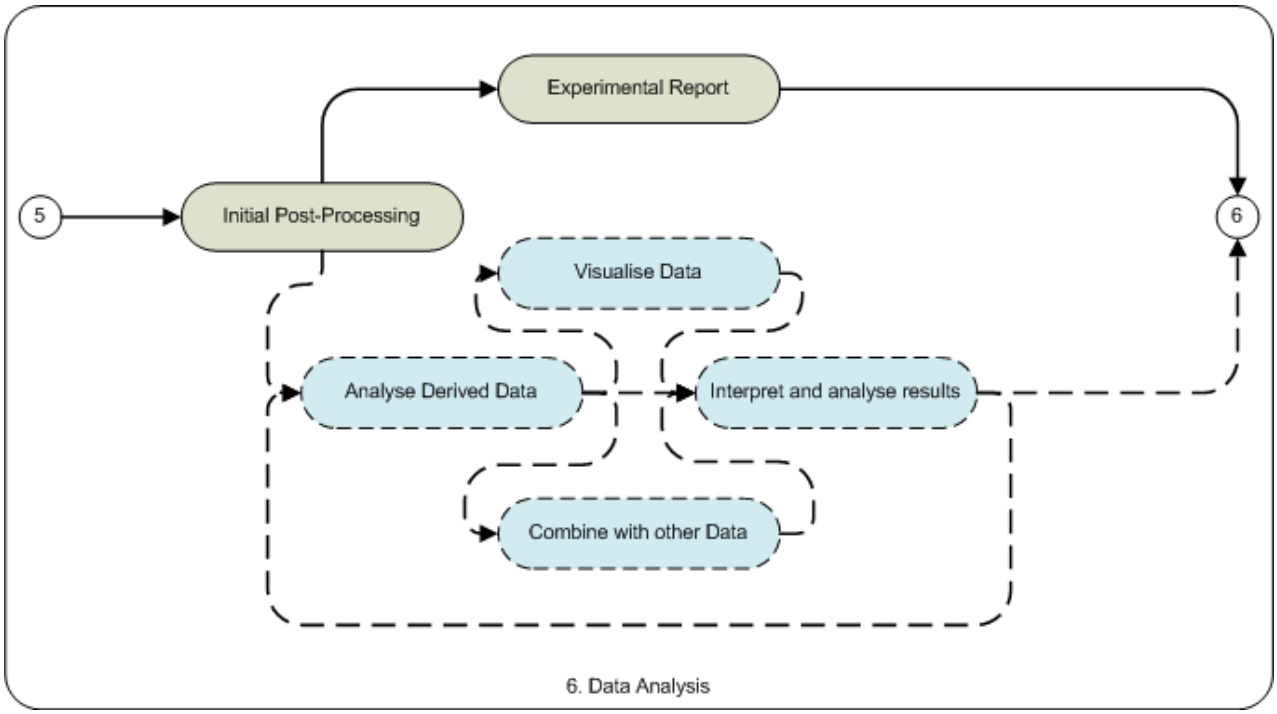
Description

The experimental scientist takes the results of the experiments (the “raw data”), and carries out a number of analysis steps. Typically, the data arising from the instruments is in terms of counts of particles at particular frequencies or angles. This needs highly specialized interpretation to derive the required end result, typically a “picture” of a molecular structure, or a 3-D image of a nano-structure. Further the interpretation needs to take place in the context of calibration or reference data, which provides a back ground in which to assess the numbers. Thus the use of highly specialized analysis software is required This may be provided by the facility itself, especially in the early stages of this process, where standard reductions are taken, or else within the experimenters research lab, on their own computers where may apply their own models and theories. This may take place over a period of months or years while the investigators derive the desired quality of result.

Sub-processes

The analysis process is typically very unpredictable, and much of it takes place within the user scientists’ institution and under their control; again much of the intellectual input of the scientists is involved in this part of the process, and the services of the facility staff have limited input. Here we give an outline of the general types of stages which are carried out in this stage of the scientific

process.



- **Initial Post-Processing:** Initial post-processing of raw data may be relatively standardized, generating processed data. For example a "reduced" data set may be generated which is the result of comparing raw with calibration data and with background noise removed. This stage is often undertaken in the facility using standardized methods and software.
- **Analyse Derived Data:** further analysis steps are undertaken by applying analysis software packages to the data to extract particular features or characteristics, or fit it to a model, for example to derive a molecular structure.
- **Visualise Data:** data is transformed into a graphical form which can be visualized and explored to provide a communication mechanism to the user scientists and more widely.
- **Combine with other data:** the data is merged or compared with other data, taken from other instruments, or from modelling and simulations.
- **Interpret and analyse results:** the results are assessed by the scientific team to determine whether the results gained so far are scientifically significant enough to warrant publication. If not, further analysis steps may be required.
- **Experimental Report:** At some point after the experimental data has been taken, the experimental team are requested to produce an experimental report on the results of the use of the facility, which should be lodged with the facility.

We discuss the factors involved in this stage further in Section 3.

Actors

- Experimental Team: directly involved in the derivation of analysed results from the collected data.

- Instrument scientist: is likely to be involved giving scientific advice and input on how to proceed with the interpretation and analysis of the data.
- User office: accepting the experimental report.

Information Systems

- Data storage systems,
- User office systems
- Analysis software packages,
- Visualisation systems

Data Types

- Processed and Derived data sets
- Graphical information for visualisation.
- Software code

Metadata Types

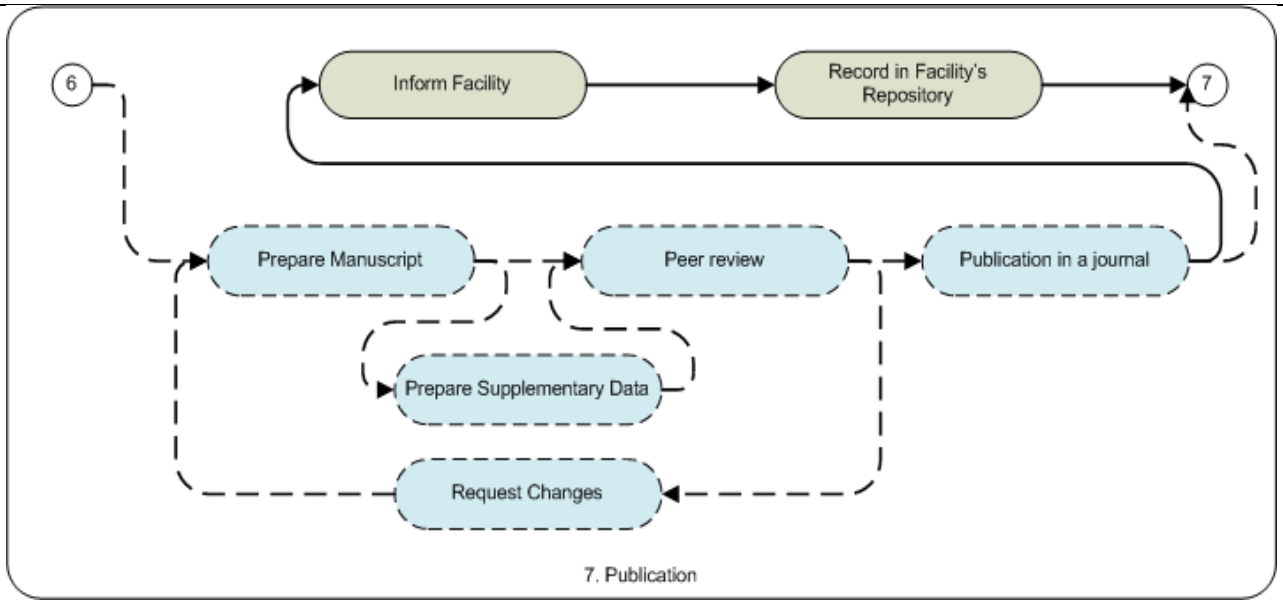
- User identity,
- Data formats,
- Data set information,
- File identifiers
- Instrument parameters
- Calibration information
- Software package information,
- Dependence tracking and workflow

2.3.7 Publication

Description

A suitable scientific result having been derived from the data collected, then the scientist will typically report the results with journal articles or other scholarly publications. The facility would usually like to be acknowledged within the article and also informed of its publication, so that it can record the value of the science derived from the use of its facilities.

Sub-Processes



This would be a standard publication process, which would typically involve at least the following sub-stages:

- **Prepare manuscript for publication:** the experimental team present the significant results in the form of an article for publication in a journal
- **Prepare supplementary data:** a data package of resultant (final analysed) data supporting the result is prepared and submitted with the paper
- **Peer review:** the paper is submitted to journal and subject to peer review, which makes a decision as to whether it is of acceptable quality.
- **Request Changes:** the review may request changes for revision (or reject the paper), leading to a likely revision of the paper and a resubmission (possibly to another journal).
- **Publication in a journal:** the article appears in a journal
- **Inform Facility:** the facility's user office is informed of the paper and records it as an output of the proposal.
- **Record in facility's library:** the facility library enters a record of the publication in the institutional repository, taking a copy if appropriate.

Again, much of the work in this stage involves the experimental team at their home institutions and does not involve facility's support staff directly.

Actors

- Experimental Team: will prepared papers
- Instrument scientist: often involved in writing the paper as an author
- User Office: record the association of a paper with an experiment
- Library: lodge a metadata record and is appropriate a copy of the paper

Information Systems

- User office systems
- Research Output tracking systems
- Library systems
- Institutional repository

Data Types • The journal article • Supplementary data
Metadata Types • User Identity • Proposal information • Publication information • Supplementary data information

2.4 Approaches to Provenance

The present data cataloguing systems within facilities only support cataloguing and accessing the raw data produced by the facility. As we can see in section 2.3, it is in the early and mid-stages of the experimental process, up to the post-processing of data, where a facility can exercise a good deal of control within its own staff and information systems. After that point, the data derived from subsequent scientific analysis is managed locally by the scientist carrying out the analysis at the facility or in their home institution. This is on an ad hoc basis, and these intermediary derived data sets are not archived for other purposes. Thus the support for tracking derived data products is partial (see Section 3 for a detailed discussion). In order to improve the support offered by the facilities the data management infrastructure needs to be extended, and in particular the facilities information model needs to cover these aspects of the process to support access to the derived data produced during analysis, and the provenance of data supporting the final publication to be traced through the stages of analysis to the raw data.

Bio-scientists have used workflow tools to capture and automate the flow of analyses and the production of derived data for many years [e.g. Oinn et. al. 2004] and can now automatically run many computational workflows. In other structural sciences, such as chemistry and earth sciences, the management of derived data is less mature, workflows are not standardised and can less readily be automatically enacted. Rather the data needs to be captured as the analysis proceeds so that scientists do not lose track of what has been done. A data management solution is required to capture the data traces that are generated during analysis, with the aim of making the methodologies used by one group of researchers available to others.

Further, the accurate recording of the process so that results can be replicated is essential to the scientific method. However, when data are collected from large facilities, the expense of operating the facility means that the raw data collection effectively cannot be repeated. Therefore tests to replicate results may have to come from re-analysis of raw data as much as repetition of the data capture in experiments.

Facilities may not consider that extensive support within this area is their prime responsibility, nevertheless there are advantages in offering some support in this area, particularly in managing early stage analysis undertaken at the facility, which is often systematic or automatable, and thus an extension of good data management practise can offer systematic tracking of derived data at relatively low cost. Further, facilities are increasingly offering “express services” where more routine experimental analyses can be undertaken by the facility on receipt of a sample without the inter-

vention of the user experimental team, which only receives the resulting data products. In this latter case, good derived data management is essential to ensure a quality result is delivered.

In order to provide support for the analysis undertaken by the experimental scientists; to permit the tracing of the provenance of published data; and to allow access to derived data for secondary analysis, it is necessary to extend the current information model to account for derived data and to record the analysis process sufficiently for the needs of each of these use cases. In terms of data provenance the current information model approach identifies the source provenance of the resultant data product, but it needs to be extended to describe the transformation provenance as well [Glavic and Dittrich 2007].

3 An example of the Lifecycle in Practice

In this section we briefly describe a specific example of (part of) an experimental lifecycle. This is the result of work previous to PaN-data originally undertaken within the I2S2 project[4] [Yang et. al. 2011]; however a summary of the work is included here as an illustration of the complexity of the scientific lifecycle associated with facilities science, and the motivation for further discussion.

The example data analysis pipeline covers the stages from the raw data collection at a facility to the final scientific findings suitable for publication. Along the pipeline, three concepts, raw, derived, and resultant data, are often used to differentiate the roles of data in different stages of the analysis and to capture the temporal nature of the processes involved. Raw data are the data acquired directly from the instrument hosted by a facility, in the format support by the detector. Derived data are the result of processing (raw or derived) data by one or more computer programs. Resultant data are the final results of an analysis, for example, the structure and dynamics of a new material being studied in an experiment.

The case study in question aimed to determine the structure of atoms using the neutron diffraction[5] provided by the GEM instrument[6] located at the ISIS neutron and muon source. The analysis workflow for this experiment involves computationally intensive programs, and demanding human oriented activities that require significant experience and knowledge to direct the programs.

In practice, it can take months from the point that a scientist collects the raw data at the facility to the point where the resultant data are obtained. The workflow has data correction process using a set of programs to correct the raw data obtained from the instruments (e.g. to identify the data resulting from malfunctioning detectors, or remove the “background signal”), though this represents only a small part of the respective workflow.

[4] Integrated Infrastructure for Structural Science (I2S2), UK JISC sponsored project, 2009-11 between Universities of Bath, Southampton, and Cambridge, STFC, and Charles Beagrie Ltd.. Example courtesy of Prof. Martin Dove, University of Cambridge (now QMUL).

[5] http://www.isis.stfc.ac.uk/instruments/neutron-diffraction2593.html

[6] http://www.isis.stfc.ac.uk/instruments/gem/gem2467.html

3.1 Data Analysis

Data analysis is the crucial step transforming raw data into research findings. In a neutron experiment, the objective of the analysis is to determine the structure or dynamics of materials under controlled conditions of temperature and pressure.

Figure 2 illustrates a typical process for analysing raw data generated from the GEM instrument using Reverse Monte Carlo (RMC) based modelling [Yang 2010]. The RMC method is probabilistic, which means that a) it can only deliver an approximated answer and b) in theory, there is always scope to improve the results obtained earlier using the same method. In the figure, rectangles represent the programs used for the analysis; rounded rectangles without shadow represent the data files generated by computer programs; rounded rectangles with shadow represent data files hand-written by scientists as inputs to the programs; ovals represent human inputs from scientists to drive the programs; solid lined arrows represent the information flow from files to programs, from programs to files, or from human to programs; and the dashed lined arrows are included to highlight the human oriented nature of these programs demanding significant expertise. This is an iterative process that takes considerable human effort.

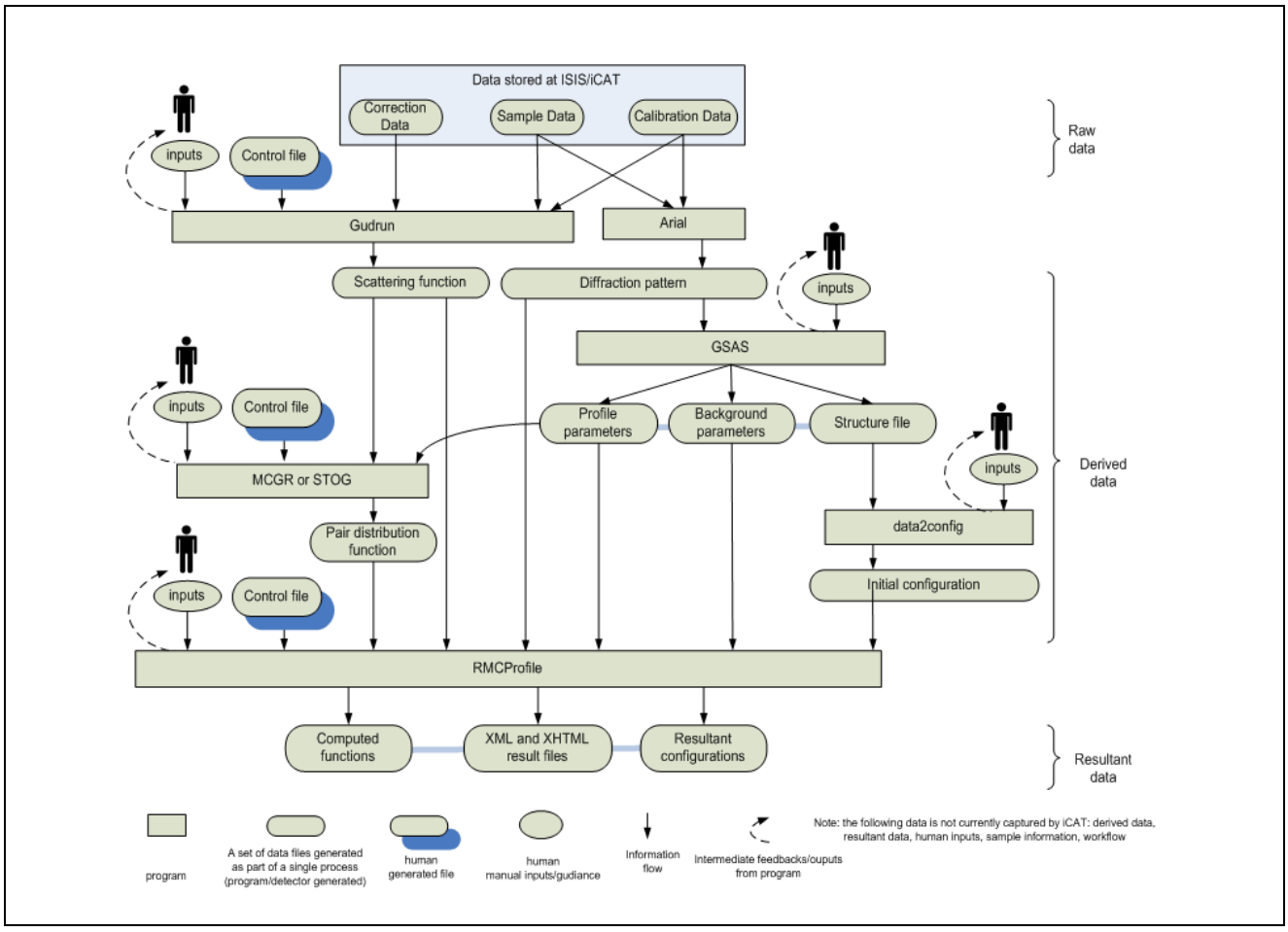


Figure 2: The RMC data analysis flow diagram

3.2 Data reduction

Three types of raw data are input into the data analysis pipeline: sample, correction, and calibration data. They are first subject to a data reduction process which is facilitated by two programs: Gudrun, a Fortran program with a Java GUI, and Ariel, a IDL program. The outputs from Gudrun[7] are a set of scattering functions, one for each bank of detectors. For Ariel[8], the outputs are a set of diffraction patterns, again, one per bank of detectors. With Gudrun, the human has to subtract any noise in the data going from scattering function to pair distribution function (through the MCGR or STOG program). Noise can arise from several sources, e.g. errors in the program, or noise due to the statistics on the data. In other words, when the other programs use the derived data generated by Gudrun, human expertise is required to steer the way the data is used.

3.3 Initial structural model generation

The next step is the process of generating the initial configuration of the structure model that will be used as the input to the rest of the RMC workflow. This step requires three programs (i.e. GSAS, MCGR or STOG, and data2config) to transform the reduced data into structure models that best fit the experimental data. To do this requires determining the structural parameters (e.g. atom positions), illustrated as the sets of data files under GSAS, for all the crystalline phases present, which are: profile parameters, background parameters, and (initial) structure file.

Most neutron and synchrotron experiments use the Rietveld regression analysis method to refine crystal structures. Rietveld analysis, implemented in GSAS, is performed to determine the structural parameters as well as to fit the crystal structure to the diffraction patterns using regression methods. Like all regression methods, it needs to be steered to prevent it following a byway. Some values in the pair distribution functions produced from MCGR or STOG are compared with their counterparts in the scattering functions to ensure that they are consistent. If they are not, the scientist repeats the analysis.

The data2config program takes the configurations generated from GSAS, or from crystal structure databases to determine the configuration size of the initial structure model.

3.4 Model fitting

All the derived data generated up to this point represents an initial configuration of the atoms, random or crystalline, which is fed into the RMCProfile [Tucker et. al. 2007] program implementing the RMC method to refine models of matter that are mostly consistent with experimental data. It is the final step in the analysis process to search for a set of parameters that can best describe experimental data given a defined scope of the search space and computational capacity. This is a compute-intensive activity which is likely to take several days of computer time. It is also a human-oriented activity because human inputs are required to “steer" the refinement of the model.

[7] http://www.isis.rl.ac.uk/disordered/Manuals/gudrun/gudrun_GEM.htm
[8] http://www.isis.stfc.ac.uk/instruments/osiris/data-analysis/ariel-manual9033.pdf

3.5 Discussion

The scientific process under consideration passes through the main phases of sample preparation, raw data collection, data analysis and result gathering. The overall data analysis process described above passes through the three phases of data reduction, initial structural model generation, and model fitting. This hierarchical structure is common to the different processes analysed. However, as the detailed example above illustrates, within each of these phases there are many different programs involved (with potentially different versions), with varying numbers of input and output objects. Because the analysis method is probabilistic, there is always scope for further improvements to the results so variations on the analysis can always be undertaken.

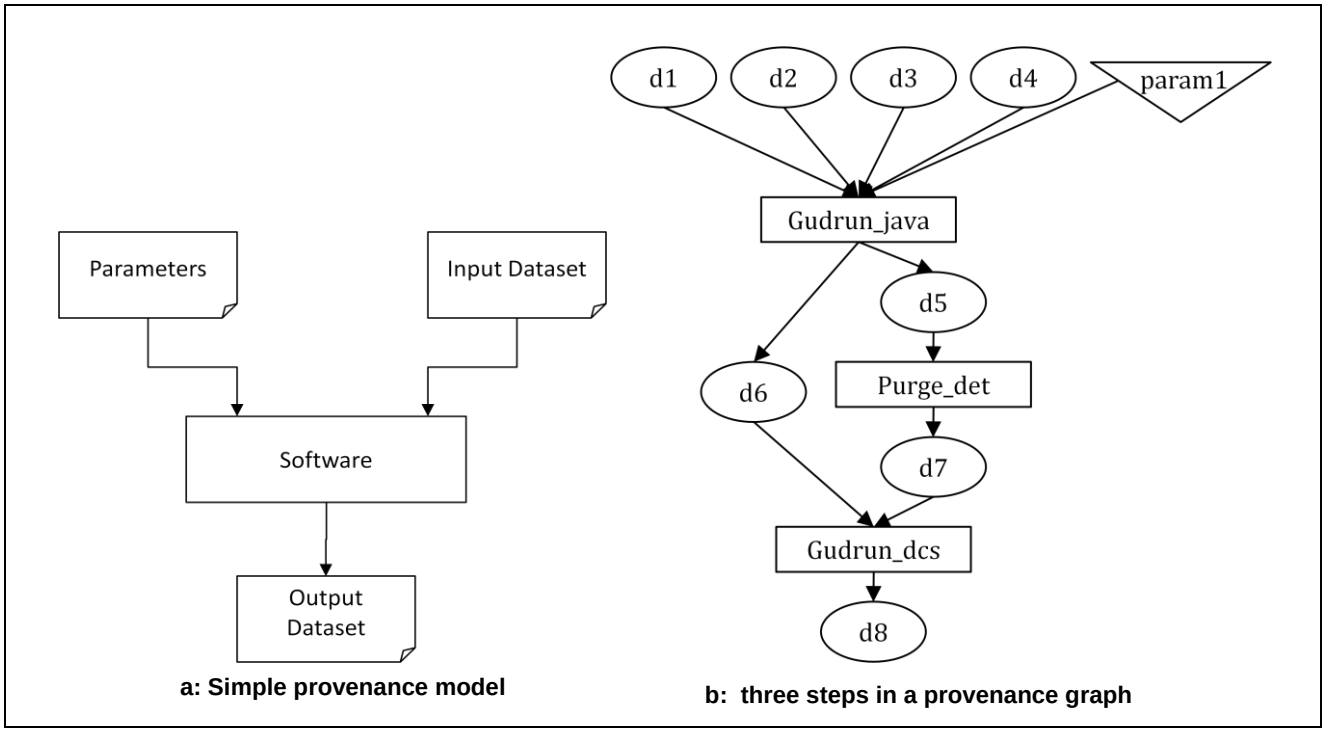
Throughout the analysis, many of the intermediate results are useful both for the scientists who perform the original experiment and others in the scientific community. The investigators or others can, for example: use them for reference; revisit them when better resources (more powerful computers, better analysis methods, programs or algorithms) are available; and revise them when better knowledge about the program behaviours are available. The scientists consulted are thus not only motivated to publish their final results but also the raw and derived data generated along the analysis flow. This is especially true for new analysis methodologies, such as the RMC method discussed here which is a relatively new method in the neutron scattering community which those who use it wish to have accepted more widely. In this case, scientists are highly motivated to publish the entire data trail along the analysis pipeline and publicise the methodology that is used to derive the resultant data. Making their data available potentially can lead to: more citations to their published papers and results; awareness and adoption of their methodology; and the discovery of better atomic models built on the models they have derived. Data archiving is also of interest to the facilities operators because of the potential of derived data reuse by other researchers who would add more value to the initial experimental time.

Thus in the I2S2 case study, a prototype was designed to capture the analysis steps via a simple provenance relationship relating: the *Input data sets* of source data together with an user modified parameters; a *SoftwareExecution***,** representing the execution of a particular instance of a software package; and *Output data sets* as the resulting data output from the particular software execution (Figure 3a). A modified version of the ICAT software catalogue was developed to capture this relationship, so that the provenance dependencies could be capture and the relationship between final resultant data and raw data audited. Thus provenance graphs can be represented as in (Figure 3b).

This approach forms a simple foundation for capturing provenance through an analysis process. However, the approach also raised issues on how to pragmatically support this approach. Some core issues were:

1. **Managing the exponential explosion of dependencies**. Even a simple step could when represented in detail contain a large number of dependencies, as illustrated in Figure 4. When such dependencies are captured across the whole length of the analysis process, and including alternative paths and parallel analysis attempts, the whole graph soon becomes very large and difficult to manage, becoming difficult to recognize the valuable dependencies

2. **Data volumes.** In a general approach, for each pathway a large number of data files may need storing, leading to a requirement for a potentially large amount of storage This is perhaps less of an issue for the end scientist, as the user scientist would typically keep multiple sets of analysed data, and capturing the provenance graph offers an opportunity to effectively manage the data so that previous analysis attempts can be found with their context and retried. Nevertheless, the open ended nature of this process would make planning storage capacity difficult for a data management service supporting provenance



a: Simple provenance model

b: three steps in a provenance graph

Figure 3: Representing provenance in the GEM example

3. **Identification of valuable data.** This approach in theory offers the capability of capturing all paths undertaken in the analysis process. In the gaining of a specific end result, a critical pathway could be reconstructed through the dependency graph to encapsulate the key decisions. However, many pathways undertaken during an extended exploratory process of analysis are likely to be erroneous, dead ends with no real gain, or representing decisions which were not followed up and have no meaning and have little real value for the future auditing, retracing and potential reuse. There are likely to be a smaller number of key decision points where valuable advances have been gained in the analysis, and alternative paths could be taken in a future re-analysis to provide new insights. Identifying the valuable paths within this large collection is therefore a difficult task, and this could lead to an obscuring of the useful data and thus make provenance information difficult to use in general.

4. **Software versioning and preservation.** A key aspect of this provenance tracing is not only to capture the dependencies between data, but also the context in which the data is processed. In particular, this means capturing information about the software packages used so that how the pathway has been constructed is visible, can be understood and validated. Further, if the analysis is to be recapitulated, then access to the software needs to be made available, so the software used should be preserved as well as the data. This is

complicated by the nature of software which is highly variable in the version and configuration (including auxiliary modules) used, a complexity which is particularly acute in scientific analysis where many software packages are written and customized by the scientists themselves (indeed this may represent much of the intellectual input of the scientist in developing novel analysis techniques), making the particular software code used at any time difficult to track and preserve.

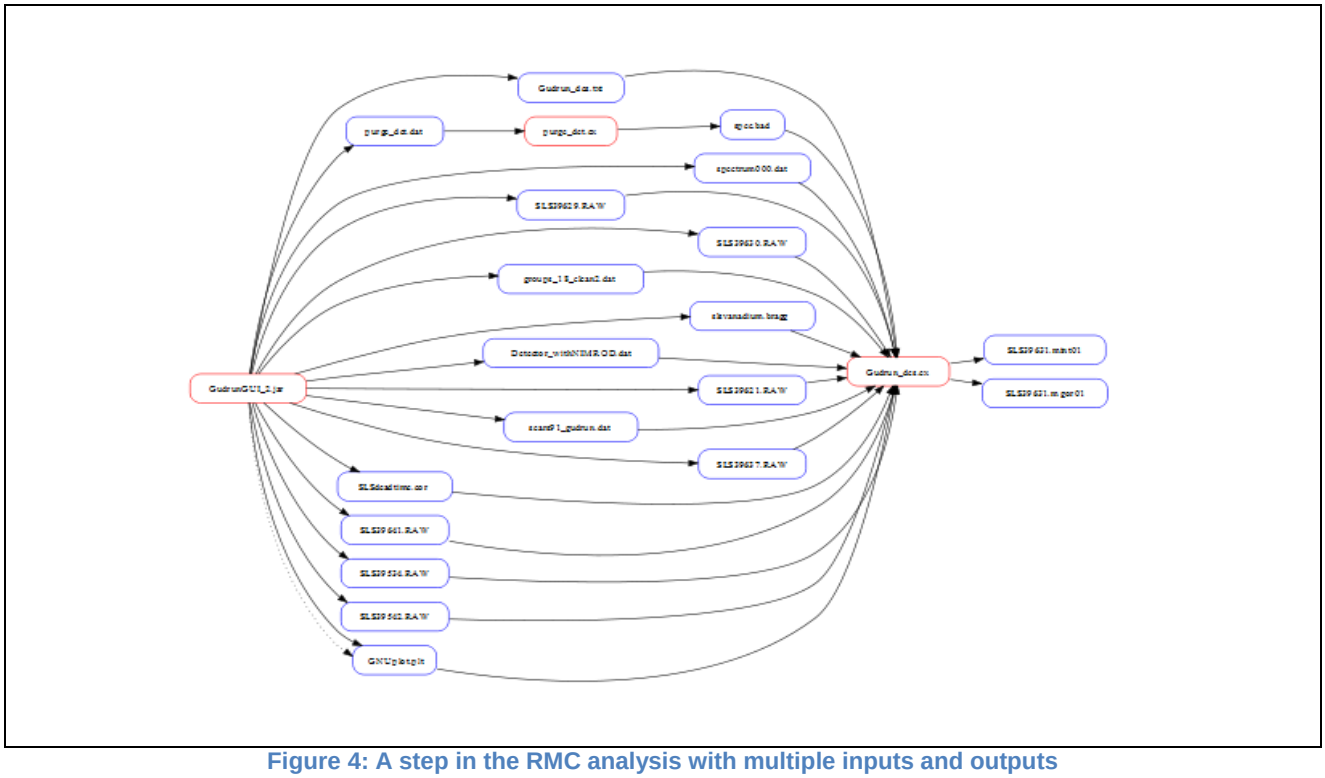


Figure 4: A step in the RMC analysis with multiple inputs and outputs

5. **Distributed analysis.** During facilities experiments the raw data is taken and stored at the facility, and some of the early stage analysis steps are frequently undertaken at the experimental facility, using software packages supported within the facility. However, user scientists' will often then take a copy of the data out of the facility for further analysis at their home institution, within their university infrastructure (including central HPC service) or using their own personal computers and laptops, taking the analysis process out of the domain and oversight of the facility's infrastructure. The user scientists may use a variety of software tools and packages for analysis and data management. This distributed analysis process makes tracing provenance particularly difficult; there is no central control over capturing the provenance trail which needs to be coordinated across a number of locations, systems and people. While linked data sharing approaches may make this tractable, it remains a difficult problem to coordinate.

6. **Role of workflow.** Some approaches to tracing provenance are based around the use of workflow management tools. This requires the description of a workflow to be designed in advance, and then enacted, with parts of the enactment potentially being automated; the provenance pathways are thus easily captured by the workflow tools. This is well suited to

"routine" scientific analysis processes, where a number of established analysis steps can be defined and executed, and reused in different analyses[9]. However, in analyses such as that in the example above, it is hard to establish a single fixed workflow; the scientist involved will often deviate from a predetermined path, try out new techniques and tools, modify software. So while parts of the process are predictable and amenable to workflow (particularly in early stage processing of raw data) this is not appropriate in general; often the stages least amenable to a predefined workflow are the scientifically most interesting.

7. **User interfaces and Integration with tools.** Recording provenance is burdensome to the user. Capturing what processes have been applied to data, which software with which parameters, and with what result forms quite a significant overhead to the busy scientist, especially in the detail required. This is information which should be captured in laboratory notebooks, but is often more ad-hoc. To make a provenance system practically feasible, it should be as non-intrusive as possible, either very easy to register those provenance steps to be recorded, in an electronic laboratory notebook system say, or by automatically capturing the provenance information, by using "provenance aware" tools, execution frameworks or rule systems which capture provenance metadata. Similarly, tools and user interfaces are needed so that provenance information can be usefully searched, explored and played back so that the benefits of capturing provenance metadata can be realized.

3.6 Conclusions on Provenance

Provenance is still an experimental area within PaN-data, with not all partners regarding it as a core part of the infrastructure, but rather within the scientific user community, and not necessarily delivering benefits which outweigh the additional costs in storage, tooling and expertise, as shown in the user survey [PaN-data-Europe D7.1]. As we have discussed above, providing a universal solution to provenance is a difficult problem, and is probably too complex and expensive at this stage.

Nevertheless, it is potentially of great value, and in scenarios where provenance can be captured and utilized effectively within the facilities data management infrastructure, and with identifiable additional cost, it can make the scientific process more efficient and lead to better science. Thus the use of provenance is scenario dependent; in this work package, we are identifying scenarios where we can apply provenance techniques and demonstrate additional value from its use. In the rest of this deliverable, we identify some initial scenarios where we can apply provenance techniques.

[9] See for example myExperiment: http://www.myexperiment.org which has developed many workflows largely in the life sciences.

4 Scenario 1: Provenance@TwinMic

Facility: Elettra synchrotron radiation facility (TwinMic beamline).

Scenario 1 is centred on the TwinMic X-ray spectro-microscope, a beamline in the synchrotron radiation facility Elettra. It combines two core modes: i) full-field imaging and ii) scanning X-ray microscope in a single instrument. It has wide range of applications including biotechnology, nano-technology, environmental science & geochemistry, clinical & medical applications, new energy sources, biomaterial, cultural heritage and archeometry.

4.1 Scientific Instrument and Technique

The TwinMic X-ray spectro-microscope is a world-wide unique instrument that combines full-field imaging with scanning X-ray microscope within a single instrument. The instrument is equipped with versatile contrast modes including absorption or brightfield imaging, differential phase and interference contrast or Zernike phase contrast - as you are used from a visible light microscope. The microscope is operated in the 400 - 2200 eV photon energy range or as equivalent 0.56 - 3 nm wavelengths. According to the energy and X-ray optics TwinMic can reach sub-100nm spatial resolution.

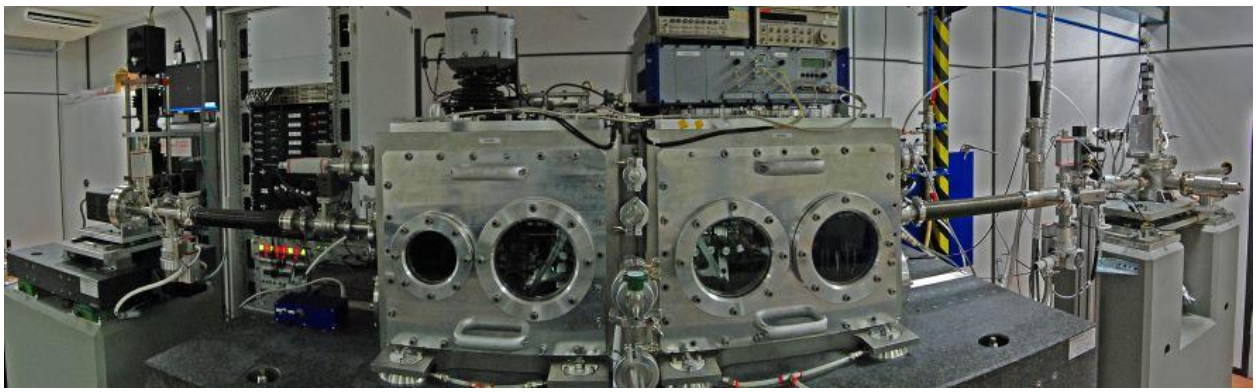


Figure 5: Part of the TwinMic Beamline at Elettra

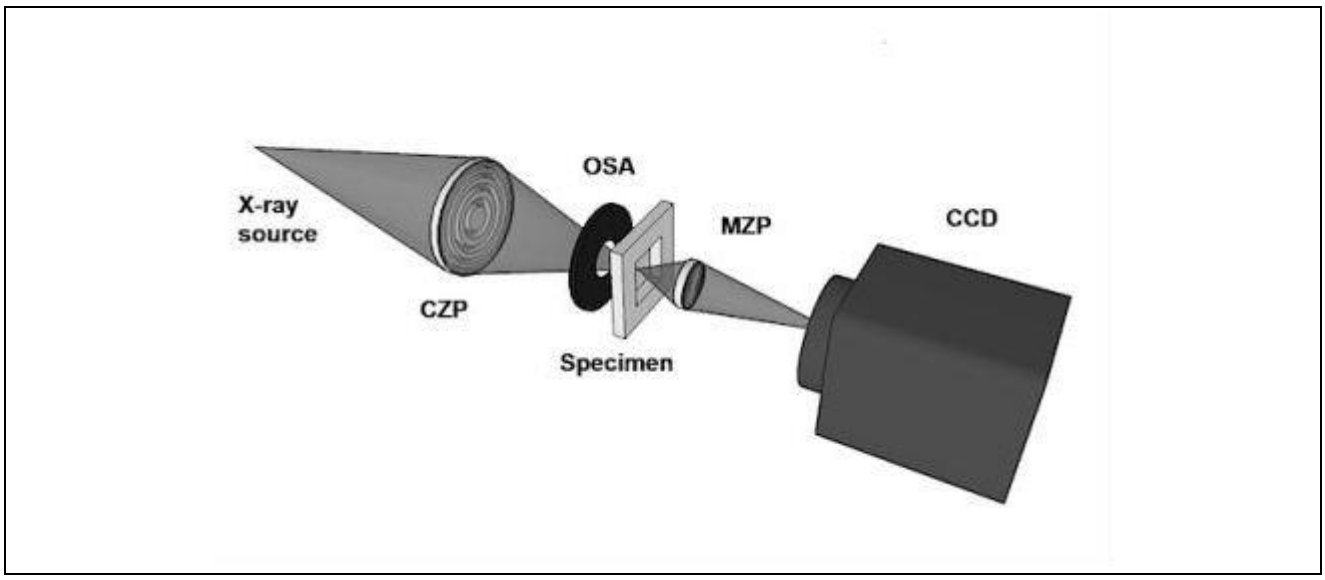


Figure 6: Outline of Full-field imaging setup in TwinMic

Full-field imaging is the X-ray analogue to a visible light microscope. A condenser illuminates the specimen and an objective lens magnifies the image of the specimen into a spatially resolving detector like a CCD camera. Since the refractive index of X-rays is slightly smaller but almost equal to unity, we cannot use refractive lenses but diffractive focusing lenses, so called zone plates. Full-field imaging is typically applied when highest lateral resolution or dynamic studies (in the second range) is required. The full-field imaging mode is limited in acquiring chemical information but we also perform X-ray absorption spectroscopy in the full-field imaging mode by across absorption edge imaging.

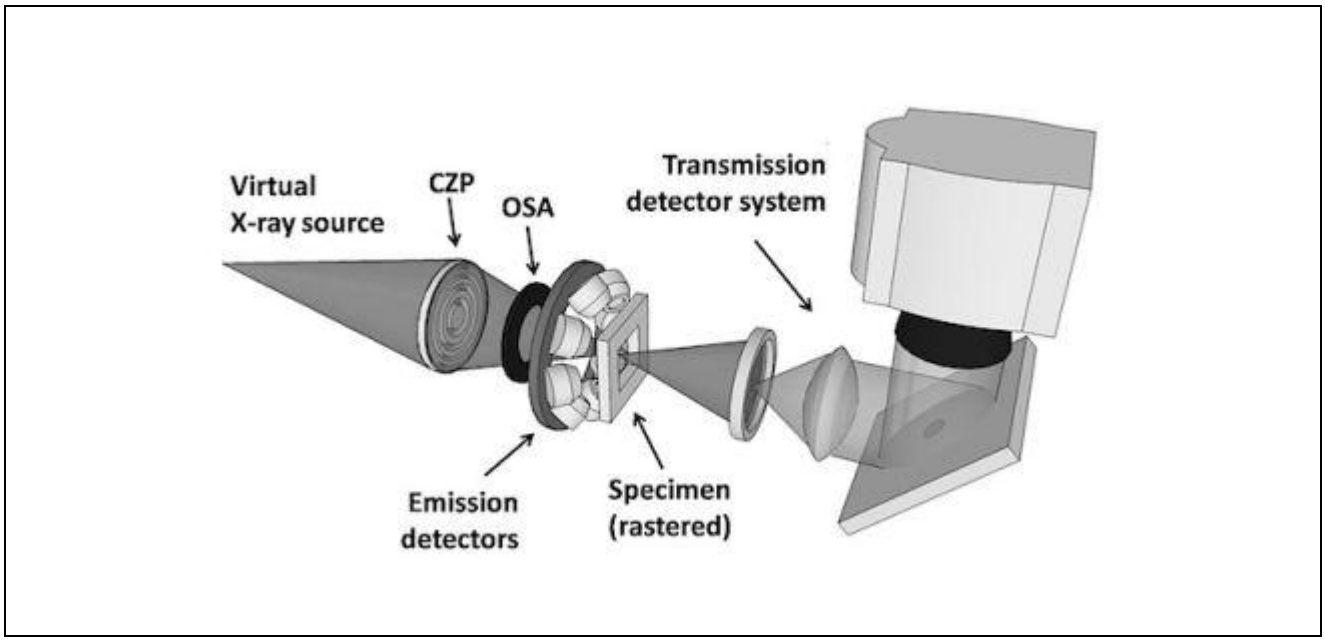


Figure 7: Outline of scanning X-ray microscopy setup in TwinMic

In scanning X-ray microscopy, a diffractive focusing lens forms a microprobe and the specimen is raster-scanned on pixel by pixel base across the microprobe. As in other scanning microscopies, this imaging mode allows simultaneous acquisition of different signals by multiple detectors (see

below). TwinMic is worldwide unique in combining transmission imaging, absorption spectroscopy and low-energy X-ray Fluorescence[10], which allows the user to analyze simultaneously the morphology and elemental or chemical distribution of your specimen with sub-micron resolution. Scanning X-ray microscopy is non-static operation mode and lateral resolution is therefore limited by the specimen movement accuracy as well as the geometrical demagnification of the X-ray light source. Fostered by newly developed SDD detectors and customized data acquisition electronics, we successfully implemented a compact multi-element SDD spectrometer in the soft x-ray SXM instrument and demonstrate for the first time XRF with submicron spatial resolution down to the C edge. The combination of sub-micron LEXRF with simultaneous acquisition of absorption and phase contrast images has proven to provide valuable insights into the organization of materials dominated by light element constituents. The major advantage of LEXRF compared to XANES is administered by simultaneous mapping of different elements without time-consuming refocusing of chromatic ZP-based lens setups operated in the entire range of 400 – 2200 eV photon energies. A quantitative analysis of LEXRF detection limits and comparison to XANES at such photon energies is under investigation and evaluation.

4.2 Scenario Description

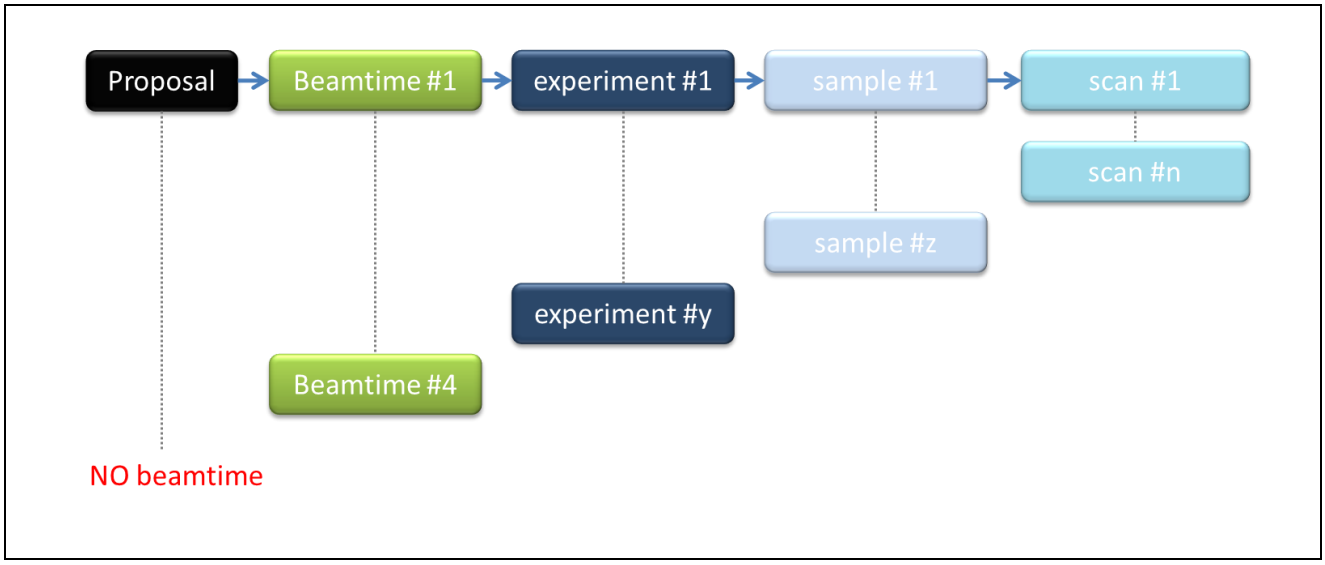


Figure 8 : Path from Beamtime proposal till individual sample scans that generate the RAW data

The backbone of the scenario connects the proposal with the data acquisition. The beamtime proposal outlines the overall project. In most cases, the proposal requests a single beamtime but it may also require more than one (i.e. long-term proposal). The proposer should state the number and type of experiments. The samples (i.e. cells) should be described in detail. A typical proposal often states the number of the required shifts accompanied with a suitable justification.

10 http://www.elettra.trieste.it/index.php?option=com_content&view=article&id=697:low-energy-x-ray-fluorescence&lang=en

After the evaluation procedure, the proposal may grant beamtime. A beamtime in TwinMic is often 9-18 shifts (3-6 days). During these days multiple experiments may be performed often taking advantages of the different modes of operation than the microscope provides.

Each experiment may involve different samples of different composition, type and preparation. These samples are often scanned/examined one or more times (i.e. different energy setup, different areas, etc.). Each scan will result in new data. The data at this stage are what the TwinMic scenario considers as RAW. Metadata at this stage are mostly information from the instrument/control system and the proposal.

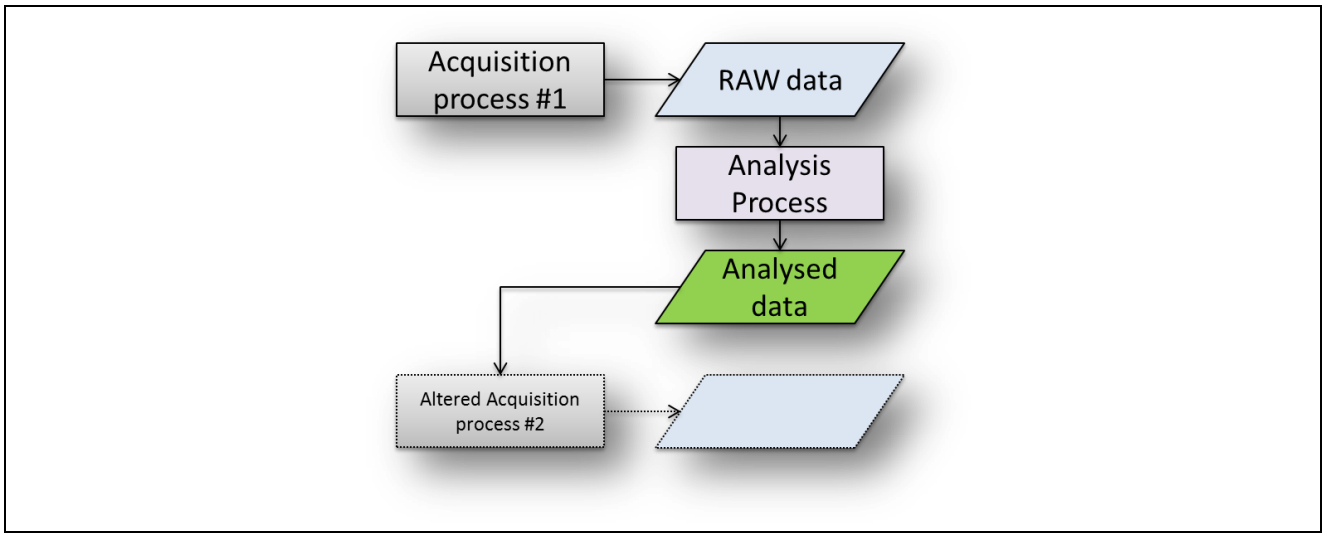


Figure 9: A series of data acquisitions that are depended to the results of the preceded ones.

The analysis and post-processing stages often take place during the data acquisition. The analysed data may alter the subsequent acquisition strategies and scans (i.e. failing at identifying a chemical element may require change of energy or sample). The systems, procedures and workflows that are already in place and support the above mentioned scenario start with the Virtual User Office (VUO) that provides the expected functionality of an advanced electronic user office platform. The main proposer needs to be a register user and all the beamtime proposal details are registered in the system. Some of this information (i.e. abstract of the proposal, sample information) may be harvested as metadata at a later stage.

An experiment may involve multiple modes and techniques as described in a later section. The two main options are i) Full field and ii) Scanning Transmission X-ray Microscopy (STXM). Each mode (i.e. STXM) has multiple techniques like X-ray Fluorescence (XRF) and X-ray Absorption Spectroscopy (XAS). Certain experiments may try to introduce or explore new methods that are not standard options in TwinMic like Coherent Diffractive Imaging (CDI) experiments.

The produced data are stored in formats that depend on the type of experiment (i.e. XRF), instrument, and/or the requirements of the analysis software. The Full field mode mostly produces images on standard formats (multipage TIFF). For X-ray Fluorescence (XRF) scans the beamline has recently designed an HDF5 based format that takes into account the instrument's setup and the requirements of the main analysis software.

Other than generic high-level approaches to analysis (Matlab, IDL, Igor Pro, LabView), the XRF experiments rely mostly on PyMCA, Spectrarithmetics, GeoPIXE, and AXIS2000. The endstation control and frontend interface is on LabView while certain components are TANGO.

For clarity purposes we outline a specific usage scenario:

A university professor applies for beamtime with a proposal that focuses mostly on cells that need to be XRF scanned. He registers in the VUO, submits the proposal after communication with the principal beamline scientist of TwinMic. The proposal is accepted and the beamtime is allocated. The professor is accompanied by a research team of 3 other researchers who need as well to make an access request. While the experiment is performed a series of samples is scanned in TwinMic in XRF modality. The operation is controlled by the beamline scientists or from her assistants by using a LabView system. The data are stored in a network drive that can be accessed by the beamline personnel and the authorized visiting researchers. The raw data are converted in a TwinMic specific HDF5 that is compatible with the PyMCA[11] X-ray Fluorescence Toolkit of ESFR. Expert in-house personnel prepare PyMCA configuration files that will be used for the final analysis of the data. The visiting users collect the configuration files and the HDF5 for analyzing them in PyMCA. The VUO will store and information like evaluation and publications related to the beamtime.

4.3 Stages of lifecycle covered in the scenario

The stages covered in the Provenance@TwinMic scenario are in accordance to those presented in a previous section of this deliverable. Certain stages like that of [Data I/O] (Storage) may not necessarily provide all the desirable services like advanced cataloguing and data provenance tools. Other stages like the [Experiment] may be a superclass of other stages like [analysis] and [visualisation]. Finally the workflow for the stages of this scenario may not be linear and includes feedback loops (see [pre-analysis] stage).

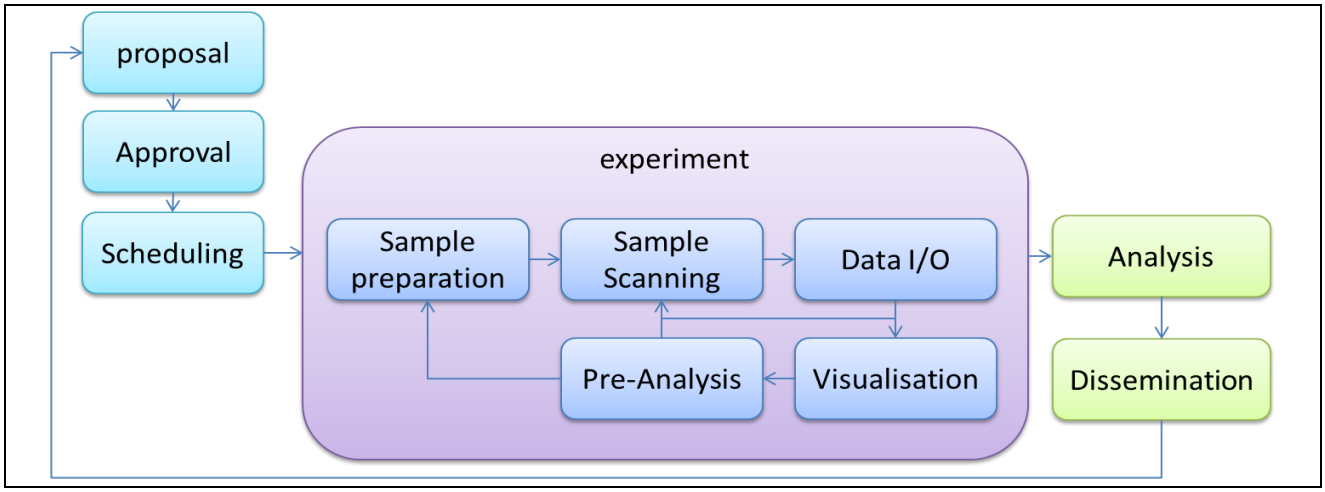


Figure 10:Stages in the Provenance@TwinMic scenario

[11] pymca.sourceforge.net

4.4 Data types

TwinMic distinguishes 4 types of data and 3 of metadata.
TwinMic data

1. RAW: the output of the endstation acquisition programs
2. Alternative RAW: converted RAW data to a lossless equivalent (i.e.NeXus) for compliance with specific software.
3. Pre-Processing supplements: data (i.e. manual peak identification) that are useful for processing the RAW – often manually generated in a preprocessing step.
4. Analysed data: processed and post-processed experimental data.

TwinMic metadata

1. Acquisition Metadata: information about the state of the setup, geometry, energy etc often acquired automatically from the control system.
2. Descriptive Metadata: information about the experiment and sample often in free text.
3. Analysis Metadata: meta-information produced in the analysis stage.

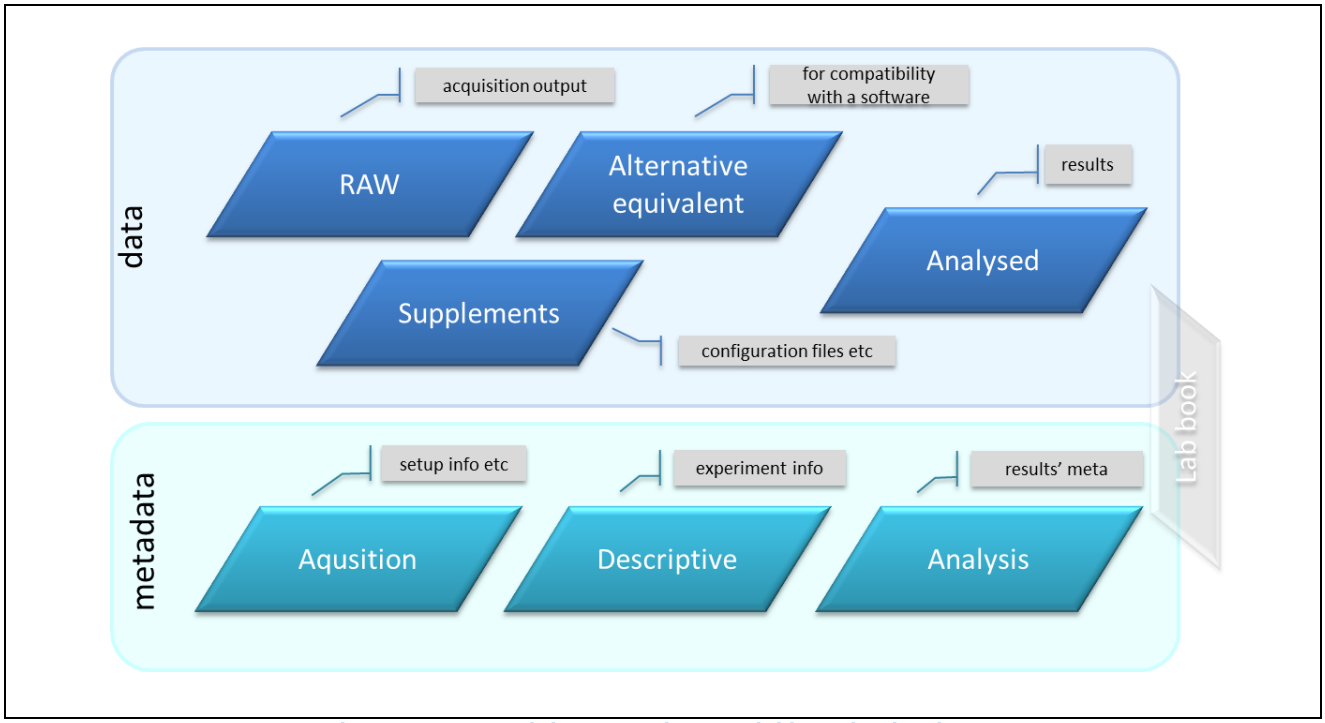


Figure 11: General data-metadata model in TwinMic-Elettra.

All the data are stored in Storage Area Network (SAN) enterprise level infrastructure. There is web-portal that permits data access for user outside the facilities network. Often users collect their data in external portable devices. Data cataloguing and provenance services and procedures are in design phase and not fully deployed. Tracking of publications is done through the VUO and the beamtime evaluation reports.

4.5 Actors involved in the scenario

There two main clusters of actors: facility (internal) and visitors (external). The core facility personnel in the scenario are the Beamline staff that includes the principal beamline scientist[12] and addi-

[12] also known as *beamline responsible* or *first beamline scientist*

tional members like second beamline scientists and post-doctoral researcher. One of them is officially associated with the proposers as their “local contact”. Additional teams of internal actors have important roles like the User Office that handles various administrative issues of the project and other on-demand teams that may need to provide ad-hoc setup and solution to the experiment (i.e. scientific computing for new algorithms, mechanical engineering for new setups etc).
The external actors are often the proposers[13]. They are often research teams with a principal investigator/scientist that is the main proposer. This person is often accompanied by a number of co-researchers that often share the beamtime shifts.

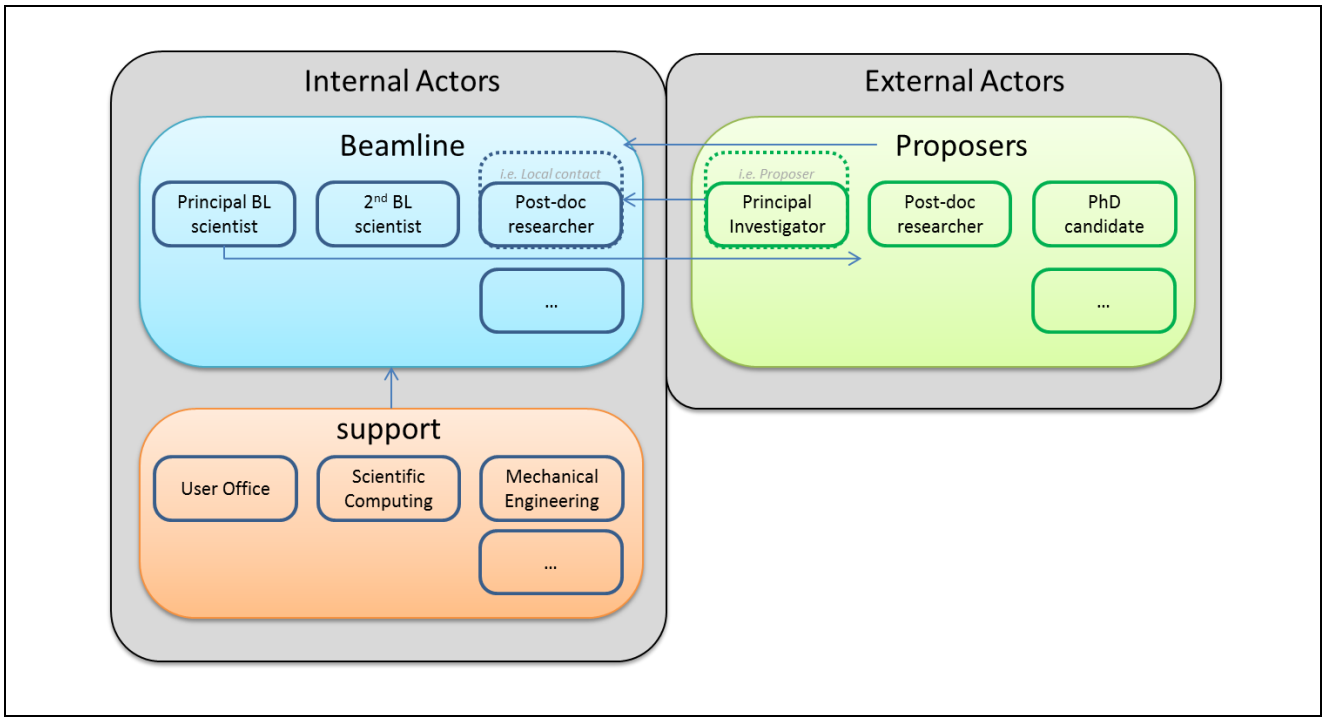


Figure 12: Internal and External actors involved in the Provenance@TwinMic scenario

4.6 Metadata requirements

TwinMic metadata:

1. Acquisition Metadata: information about the state of the setup, geometry, energy etc often acquired automatically from the control system.
2. Descriptive Metadata: information about the experiment and sample often in free text.
3. Analysis Metadata: meta-information produced in the analysis stage.

An additional type of information that could be used as metadata is in the lab-book of the beamline. Often the users are asking for copies so that they can refer to conditions and other useful notes made during the experiment. The current version is a traditional paper-based one. There are certain initiatives that aim at digitizing parts of this.

[13] even if internal user may also submit beamtime proposals

5 Scenario 2: The Smart Research Framework for SANS-2d

Facility : ISIS pulsed neutron and muon source, STFC, UK
Scientific Instrument: SANS2D
Technique: small-angle scattering
Discipline: Structural Biology, materials

5.1 Information Systems involved

ICAT	STFC's institutional data cataloguing system
ePub	STFC's institutional publication repository
LabTrove	A web-based electronic notebook system developed by the Southampton University, UK
SampleTracks	Prototype system for registering and tracking samples.

5.2 Actors

Stage	Actors	Systems	Metadata Types
1. Sample registration	Experiment team	SampleTracks	Sample name, sample parameters (composition, container size, density)
2. Setup experiment	Instrument scientist, experiment team	SampleTracks	Rack position, run duration, run type (SANS or TRANS run)
3. Data collation	Instrument scientist, experiment team	SampleTracks	File association with the samples, calibration data, and instrument data
4. Data reduction	N/A (because it is automated)	SRF Engine	Association between reduced and raw data
5. Data storage and cataloguing	ISIS computing team, data infra-structure team, in-strument scientist, experiment team	SampleTracks, LabTrove, ICAT, ISIS file store, STFC data archive	Information about experiment, reduction, and the catalogue of raw and reduced data, status of experiment (in ELN)
6. Publication	data infrastructure team, facility library, instrument scientists (add the links), ex-periment team (add the links)	ICAT, ePub	Publications, linking between data and publications

5.3 Data types and Repositories

Sample data	Sample register in the risk management system
Raw data files	File system
Metadata of raw data	ICAT
Reduced data files	File system
Metadata of reduced data	ICAT and LabTrove
Publication	ePub
Metadata of software	Software repository
Metadata of data processing	Workflow, data provenance

5.4 Scenario Description

The SRF project (aka. SRF) is a pilot project for the ISIS SANS2D instrument at the Rutherford Appleton Laboratory in the UK. As illustrated in Figure 13, SRF aims to expand the experiment process of the facility lifecycle to cover the following activities: sample registration, experiment setup, automated data tracking and collation, and automated data reduction. As the time of writing this document, it often takes several days or sometimes weeks before scientists can obtain feedbacks about their experiment. Integrating the processes into the facility lifecycle allows scientists to not only gain quick feedbacks from the live experiments, but also leverage the feedbacks to adjust the experiments accordingly.

Traditionally, for the purpose of risk monitoring, the general information about samples, such as radiation levels, are required to be logged with ISIS prior to the start of an experiment. However, the level of granularity of these information is not sufficient to identify the actual samples that are placed on the sample racks of an instrument during an experiment.

In SRF, the detailed information about samples, for example, the compound, concentration and density of the samples, the thickness of the sample containers, and the role of samples (e.g. background, direct beam, normal sample), are required to be systematically registered in the system by the investigators, prior or during an experiment. This step creates a systematic record of the samples used in an experiment.

Scientists are then required to allocate the registered samples to the sample rack positions. Each run uniquely corresponds to a sample, although a sample can be the subject of multiple runs. Scientists also specify how each experiment run is set up: this includes specifying the duration of a run and the type of a run (a SANS run or a TRANS run). This is the step of experiment setup. The output here is a set of Open Genie scripts corresponding to the selected samples, the rack setup and the run configurations for the samples.

Each experiment run in the script is assigned a unique identifier, called a run identifier, which allows the run to be traced across systems in the ISIS data infrastructure. These identifiers are fundamental to the correct operation of the SRF system. Any experiment run that is included in an Open Genie script uniquely corresponds to a nexus file logged in the ISIS journal file. Through the run identifier, the SRF services is able to automatically track and collate the file with the experiment run, hence the sample that is used in the run.

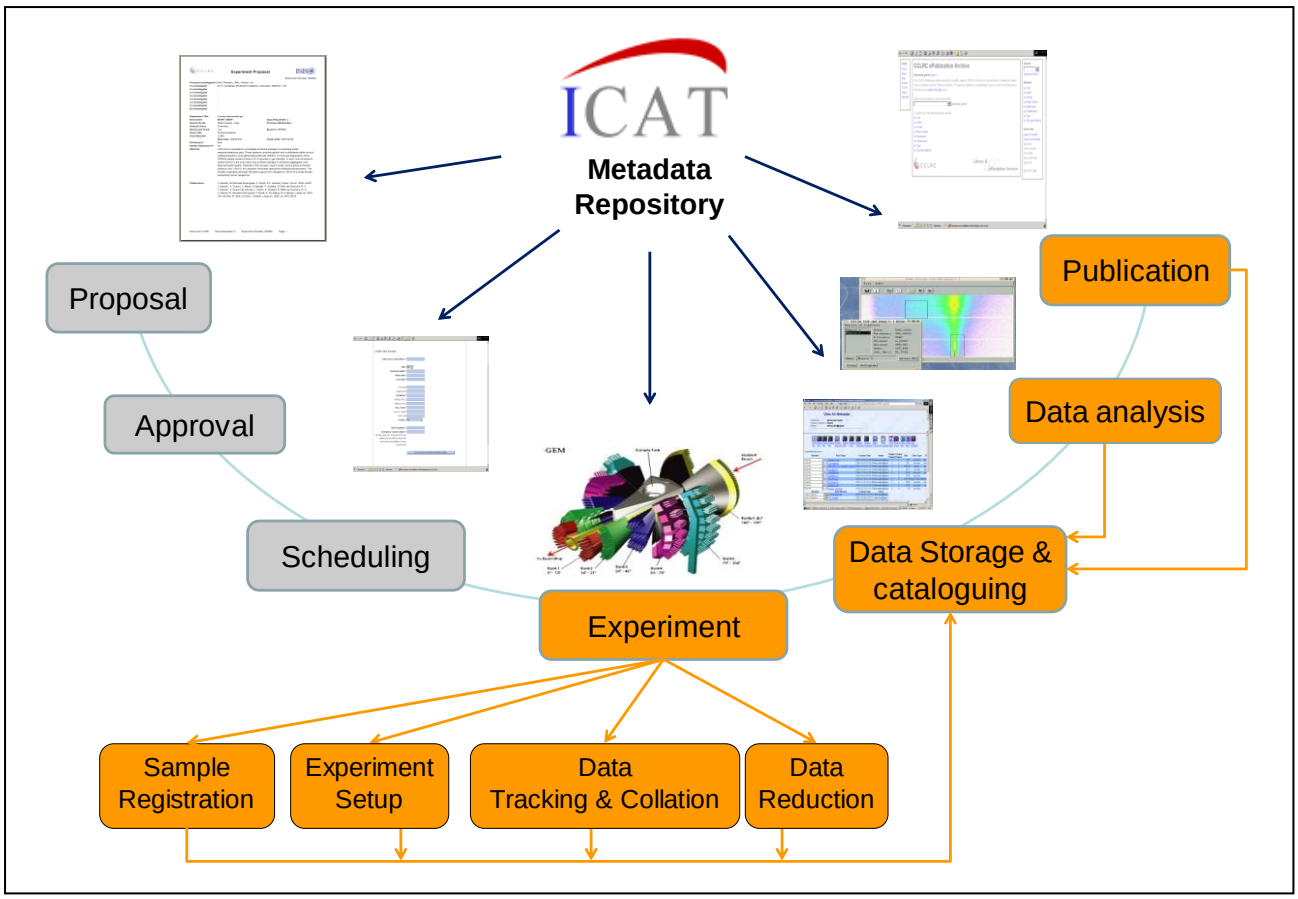


Figure 13: Processes targeted by the SRF case in the facility lifecycle: the orange boxes depict the processes targeted by the SRF case

The availability of a new data file triggers the system to check whether it is ready to perform one or more reductions. This is the automated data reduction step in the experiment process. The reduction is done via a python script which is a wrap-around of the Mantid data analysis framework.

At the end of a reduction, references to the raw data files and the reduced data files are posted to LabTrove, the electronic notebook system. Note that the raw data files are catalogued by ICAT as before. The reduced data files can, in principle, also be catalogued by ICAT as the derivative data of the raw data, however, the current implementation is not yet in place to enable this functionality. Similarly, a simple model fitting script, a wrap-around python script of the SansView framework is integrated into SRF to facilitate simple data analysis. Similarly, if needed, the analysed data can also be catalogued by ICAT.

Finally, SRF also leverages the WebTracks protocol [Crompton et. al. 2012] to enable the linking between the raw and derived datasets in ICAT and the publications in ePub, the STFC institutional repository.

6 Scenario 3: Tomography Data Processing (TDP)

Facility : Diamond Light Source Ltd., UK
Scientific Instrument: X-Ray tomography beamline
Technique: 3D tomography
Discipline: bioscience, materials, geo-science, aerospace, nuclear physics, and food science

Stage	Actors	Systems	Data types	Metadata Types
1. **Data collection**	Instrument scientist, experiment team	Data acquisition, local storage	Raw data (projection images, calibration images)	Sample name, sample parameters, experimental parameters
2. **Reconstruction**	Instrument scientist, experiment team, facility computing team[14],	local storage, archival storage, HPC resources.	reconstructed data (2d images, sinograms)	Reconstruction techniques
3. **Data Cataloguing**	Instrument scientist, experiment team, data infrastructure team	Data catalogues, archival storage	Raw and reconstructed data	Sample name, sample parameters, experimental parameters
4. **Data archiving**	facility computing team, data infrastructure team	Data catalogues, archival storage	Raw and reconstructed data	Sample name, sample parameters, experimental parameters, DOIs

6.1 Basic principles of x-ray tomography imaging

Figure 2 - (a) illustrates the basic principles of the synchrotron x-ray tomography imaging technique at Diamond. Diamond's beamline is in the so-called parallel beam configuration, where a sample is place on a rotation plate during an experiment and the x-ray beams pass through the sample in a parallel fashion.

While rotating, projections of a sample at varied angles are continuously collected by the detector of the beamline. Each **projection** corresponds to a 4,000 pixel X 2,600 pixel 16 bit grayscale TIFF image [Basham 2012]. As the time of writing this document, a total of 6,000 projections can be collected for each experiment **scan** (or run) over the duration of 30 minutes.

[14] Here, we make the distinction between facility computing team and data infrastructure team. This may not be the case with all facilities. Some may combine the roles of these teams into one. In ISIS and Diamond, the facility computing team refers to the team directly supporting the facility, whilst, the data infrastructure team provides and maintains the underlying infrastructure (data archive, publication repository, network, storage etc.).

Figure 2 - (b) and (c) show the relationship between a sample and the projections. As illustrated in (b), projection images are normally taken at equally paced angles. In order to examine the effect (or impact) of various (extreme) conditions to the behaviours and structure of a sample, images can be taken over a predefined period of time. This is depicted in (c). Within the experiment chamber, the sample can be subject to a range of experiment conditions, typically vibration, shock, temperature, or a combination of these.

Therefore, unlike traditional 2D or 3D tomography, this capability makes tomography beamlines interesting and powerful because the final results from an experiment will not only reveal the 3D model of a sample, it also allows investigators to examine how the sample evolves *over time* under various conditions, thus effectively making the model a 4D representation of the sample, which is often referred as a 4D movie of the sample.

6.2 Primary raw data and secondary raw data

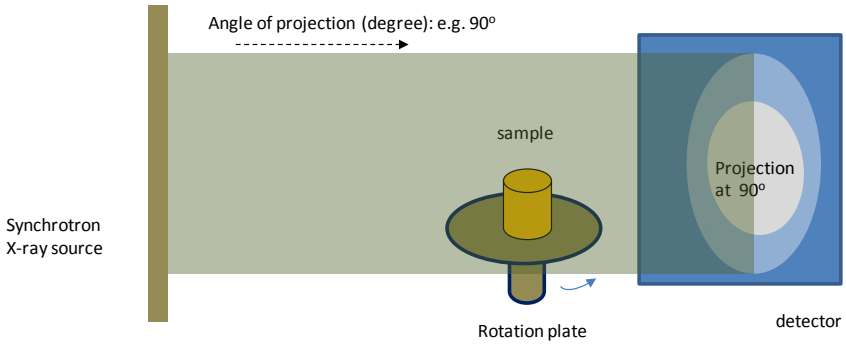
The projection images are the ***primary raw data*** from an experiment. For calibration (i.e. background correction and normalisation), additional images are also collected. This is done at the beginning and at the end of each scan. The additional images collected for calibration are called ***complementary raw data***. Two types of extra images are collected: dark- and flat-field images. These are crucial in the data reduction process for correcting and normalising the projection images.

6.3 Data processing pipeline

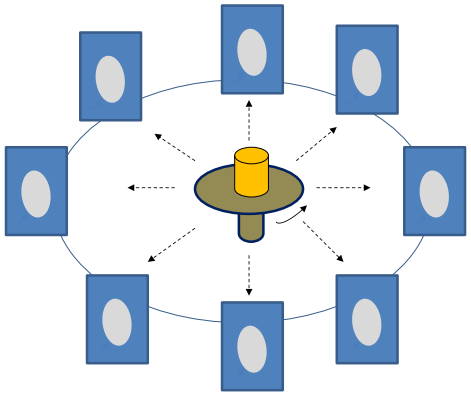
Diamond offers two tomography beamlines: i12 and i13. Both offer micro-scale tomography imaging capabilities. The TDP case study investigates the processes involved in the tomography data processing pipelines at Diamond.

Figure 3 is an overview of the data handling pipelines for the tomography data at Diamond. The rectangular boxes depict the processes involved in these pipelines. The other shapes in the diagram depict various storage entities involved in the pipelines, which includes disk, tape, and database. The left hand side of the diagram is the data processing pipeline depicting the processes from data reduction, to data reconstruction. On the right hand side of the dash line in the middle, there are two independent processes: data cataloguing and data archiving. Both processes are not only independent from each other, but also are independent from the data processing pipeline on the left.

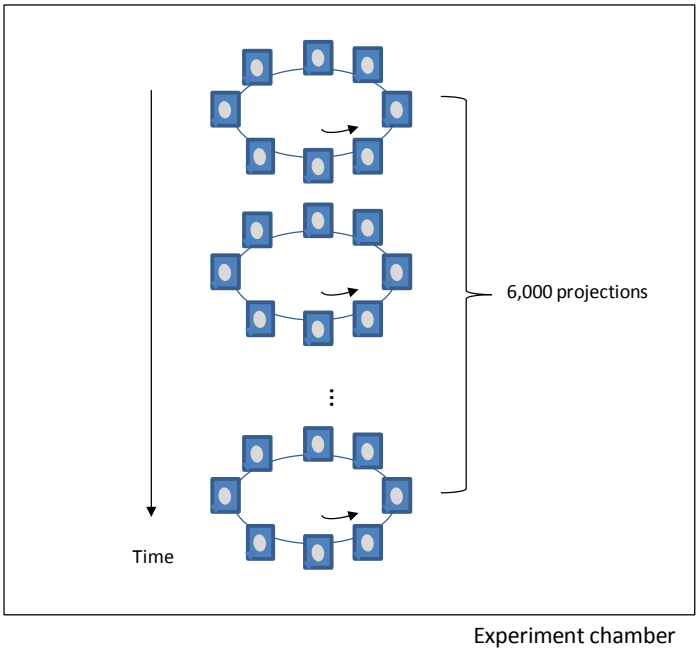
It is worth noting that the data cataloguing and archiving pipelines are not tomography beamline specific, i.e. all data from Diamond beamlines will be subject to both processes. The cataloguing process normally takes place while an experiment is still live; whilst the data archiving process takes place weekly to ensure that all the raw and processed data are catalogued and archived by the systems. Once the data is archived, they will be deleted at some point so that the disk space can be released back to active experiments.



(a) X-ray source, sample and detector



(b) sample and projections



(c): projections taken over time

Figure 14: 2 The basic principles of synchrotron X-ray tomography (a): x-ray source, sample and detector; (b) sample and projections; (c) projections collected over time.

The data processing pipeline takes about 40 minutes to go through. This pipeline is currently supported by a local 8-core[15] GPU cluster at Diamond. Both raw and processed data (including the sinograms, reconstructed images) are catalogued through the ICATingest pipeline, populating the ICAT database with basic metadata (e.g. file size, file name, directory, user's federal id) about the data.

6.4 The processes

The primary raw data can be subject to three types of processing sequentially: reduction, sinogram generation, and reconstruction. The reduction process takes the projection images and the complementary raw images to produce corrected and normalised images, which are called reduced data. Data reduction for tomography is very specific to the experiment domain (e.g. biology, archaeology, etc.).

The sinogram generation process takes the reduced data to generate sinograms. As illustrated in Figure 4, this is a fairly straight forward process, basically involving slicing *all the projections* to extract the 1 dimensional columns to produce the sinograms. The diagram illustrates that 6,000 projections, each a 4,000x2,600 px image, correspond to 4,000 sinograms. Each sinogram uniquely corresponds to an image slice of the sample.

All the sinograms are then subject to the reconstruction process, the output of which is called reconstructed data, i.e. the reconstructed image slices from the sinograms. This process takes into account the parameters to the reconstruction scripts, the characteristics of the images, and the nature of sample (e.g. fossil). These factors determine the type of the filters used during the reconstruction process. It is important to not only keep the raw data, but also the processed data along side with the parameters (e.g. filters). This is because, at the moment, the reconstructed data is the so-called "rough and ready" images, which are generated during an experiment. Scientists often want to come back to the raw images to fine tune parameters or use a different filter to produce the reconstructed data again. There are many likely causes for that. For example, scientists have a new idea of slicing the images, or years later because of the arrival of new types of techniques that allow more systematic or efficient extraction of information from the images, due to perhaps advent in the subject area or computational capabilities, or because an error is identified in the reconstructed dataset. In these cases, scientists will want to re-do the reconstruction. Finally, it is also worth to point out that reconstruction is very subjective to individual's judgements and experience. Hence, the parameters used in the reconstruction process are as important as the reconstructed images themselves.

As illustrated in Figure 4, the 4,000 sinograms produce 4,000 images slices accordingly. In the reconstruction process, the slices can be generated in parallel. At diamond, this is done with a multi-core GPU cluster which has *fast* network (10s Gbit/s) access (read+write) to the disk storing the raw and processed data.

[15] Soon, it will be upgraded to a 48-core GPU.

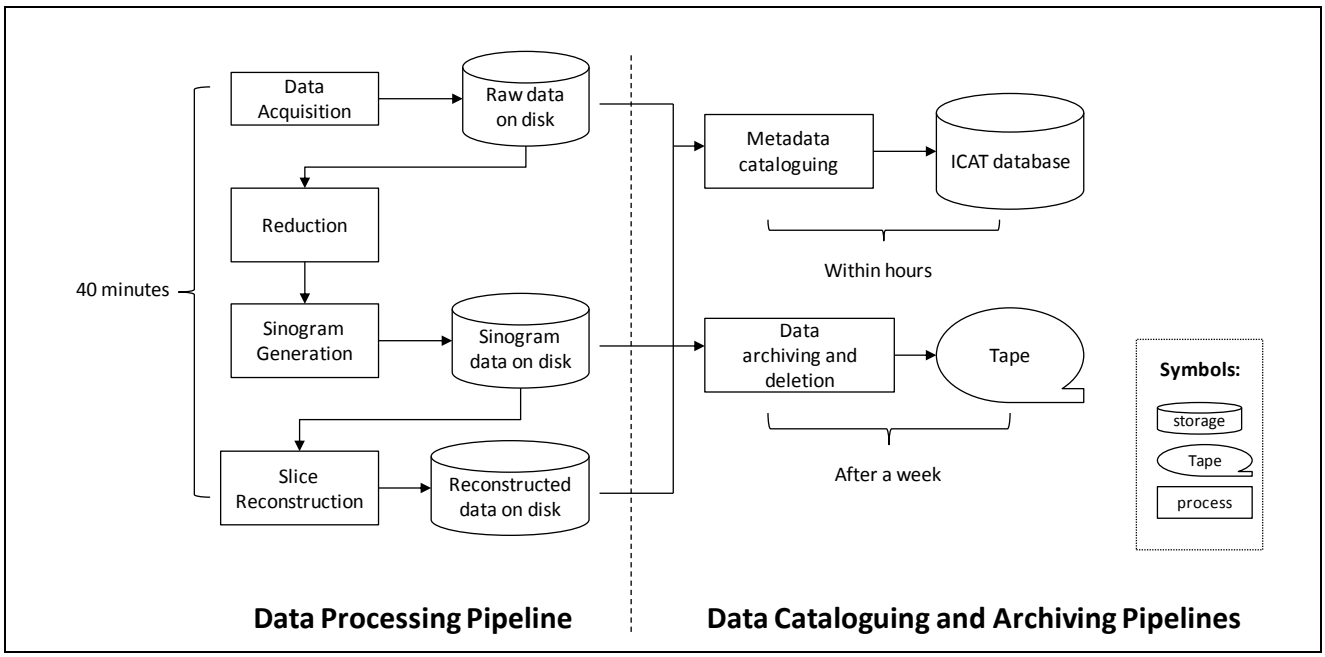


Figure 15: The Tomography Data Processing, Cataloguing, and Archiving Pipelines at Diamond

This is where the diamond data processing pipeline ends. However, this only marks the beginning of the data analysis journey that tomography users will need to go through to obtain scientific discoveries from the data they get from Diamond.

6.5 Remarks

Although the three data processing processes (i.e. reduction, sinogram generation, and reconstruction) shown in Figure 3 are presented in a linear fashion, meaning that these processes are lined up one after another sequentially, the reality is that there are many variants in this pipeline that affect the movement from one process to another. Therefore, for users, the actual pipeline of running these processes can be iterative.

6.6 Data, Metadata and Data Files

It is expected[16] that by the end of 2012, at the end of each scan, the raw data are presented as two files: a HDF5 file containing all the images (primary and complementary raw data), and a nexus file containing just the metadata. The nexus file is small, about 2MB per file; whilst the actual data in blobs is big, typically 120GB per file for every 30-minute scan. The nexus metadata file includes a pointer to the HDF5 data file. Instead of storing both metadata and the images within one big file, the attraction of storing metadata separately is that users can examine the metadata of the scan without opening the data file, which can take a long time to process. Another possibility is that a large number of metadata files can be efficiently processed without reading the actual content of the data.

[16] During the interview in June 2012, the situation is that 6,000 TIFF images from a scan are individually stored on disk in a folder designated for the raw data from that scan.

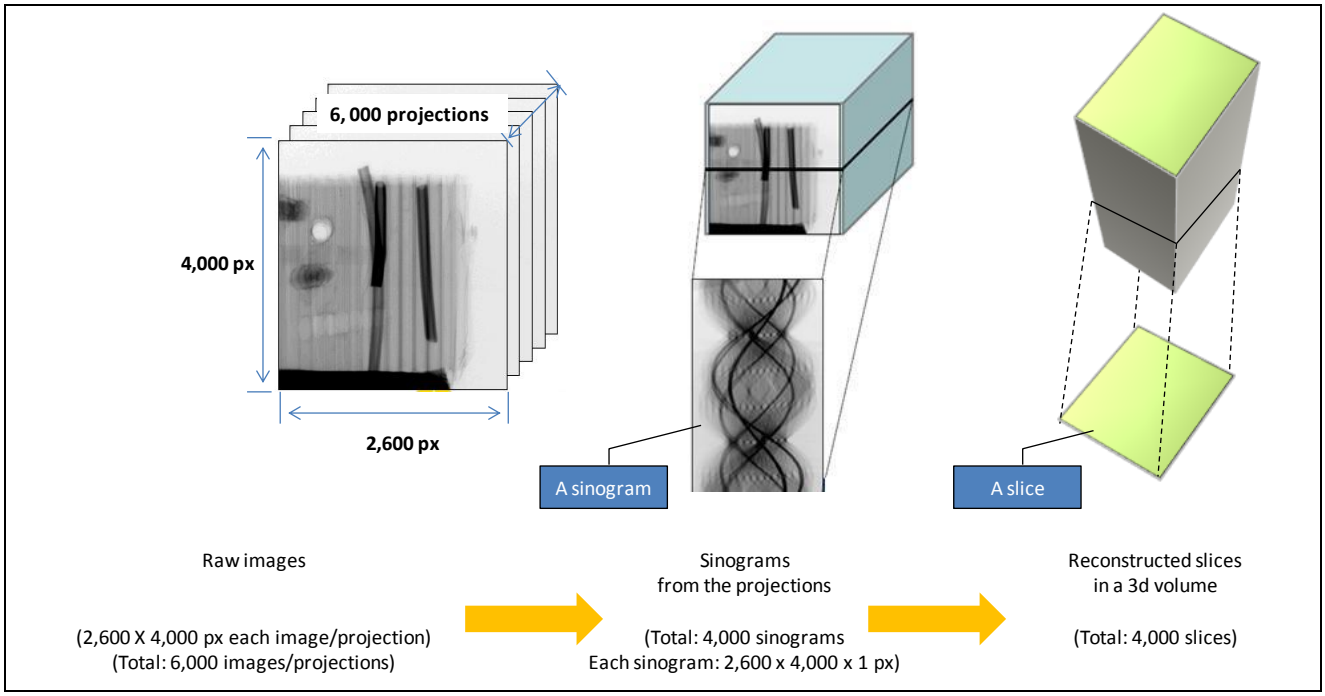


Figure 16 : Sinogram Generation and Image Slice Reconstruction (Part Image Courtesy: Dr. Mark Basham, Diamond)

The reconstructed data is also presented as two files. Each 120GB raw data file corresponds to an 80GB HDF5 file containing the reconstructed images. The nexus metadata file of the reconstructed data contains a pointer to the reconstructed HDF5 data file. The relationship between the metadata files and the data files are illustrated in Figure 5. In Figure 5, the dash line represents the relationship between the raw and reconstructed data, which is currently implicitly represented by the folder structure and naming conversion of the folders used in Diamond [Diamond-tomo 2012].

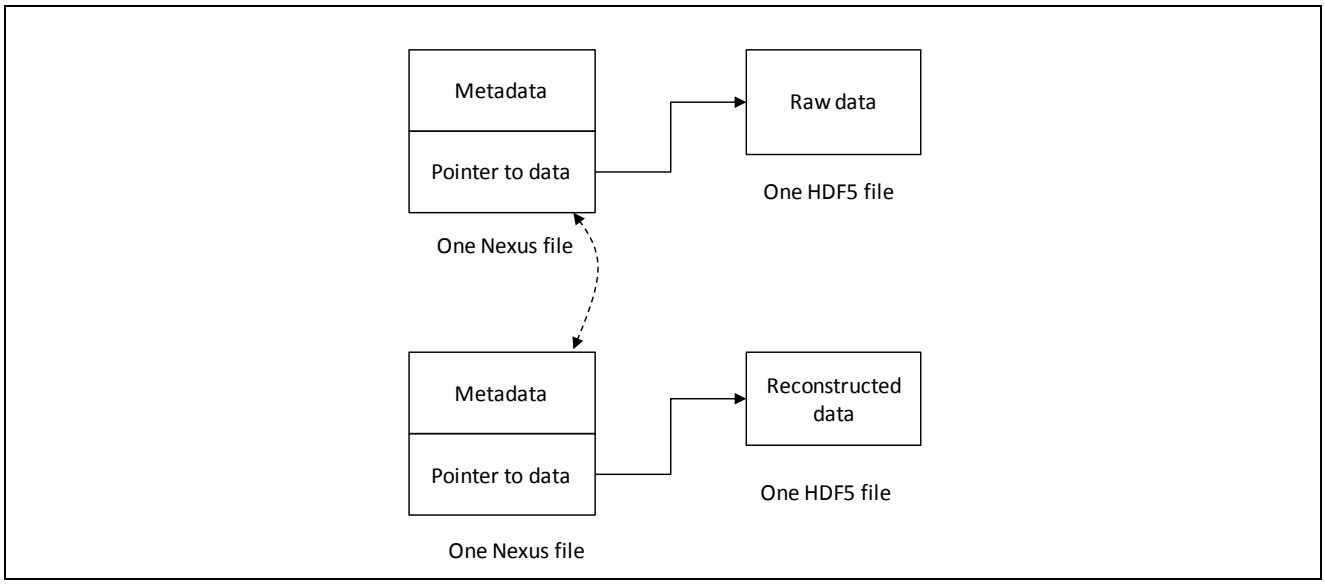


Figure 17: The Relationships between the metadata file and the data file.

7 Scenario 4: GEM Xpress (measurement-by-courier)

Facility: ISIS pulsed neutron and muon source, STFC, UK
Scientific Instrument: GEM – General Materials powder diffractometer.
Technique: Neutron powder diffraction.

7.1 Scenario Description: Powder diffraction measure-by-courier service using the Gem instrument.

Gem Xpress beam time at ISIS is mainly intended for new and infrequent users, research programmes requiring only limited and/or occasional neutron beam time and chemistry programmes operating a tight loop between synthesis and structural characterization. Fully reduced and corrected high-quality data, ready for Rietveld refinement, is provided to the user, together with example files and guidelines for refinement. The user is expected to carry out the structural analysis with minimal assistance.

To reduce the burden of running simple measurements on the instrument scientist, as much of this process is automated as possible (details below). This requires cataloging of the data at each stage to provide data provenance, should the data need to be re-analyzed manually. For Xpress measurements the user does not visit the site, so the data must be returned to them electronically – using the data catalogue. Finally, so show impact it is desirable for publications to be associated with the data.

The (formal) process begins with proposal submission using an online system (often informal discussions will have taken place between the proposer and the instrument scientist). The proposer must provide the details of the sample(s) to be measured and safety information relating to them, as illustrated in Figure 18. The instrument scientist receives these proposals as they are submitted and conducts an assessment of the feasibility of the measurement proposed and any sample or experiment hazards.

Assuming this is passed the proposal is passed to the ISIS sample safety team who review the safety of the sample and experiment and create a ‘sample risk assessment’ (SRA) for the instrument scientist. The proposer is informed of the status of their proposal, and can now request a ‘sample can’. There are two common sizes of sample can for GEM – 6mm and 8mm. The proposer requests an appropriate size and this is sent to them. They place their sample in the sample can and return it to the instrument scientist.

One day of beam time is set aside on GEM per cycle for Xpress measurements. The instrument scientist logs in to the proposal system and selects those samples he wishes to run on the allocated day.

For successful analysis of the data certain calibration files are required in addition to the collected sample data. These are: background, vanadium and an empty sample can of the same size as that containing the sample.

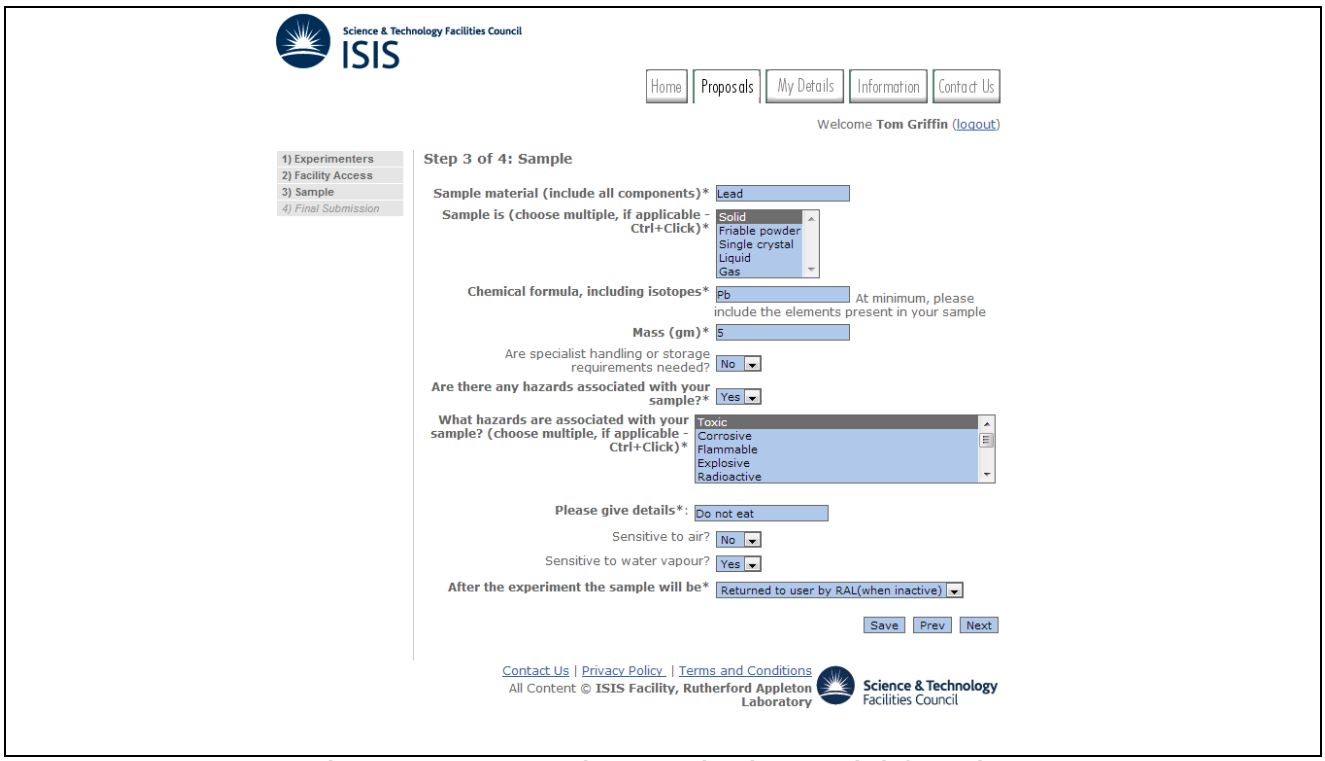


Figure 18: ISIS Proposal System showing sample information

Based on the instrument scientists sections, the system calculates an appropriate order (background and any empty can runs first) and outputs:

- A sample changer loading list for the instrument scientist.
- An instrument control script which will automate the control of the neutron instrument and sample changer, included setting identifying values in the data files.
- A data analysis script that will use these identifying values to automatically start a Mantid analysis job when the required files (calibration and sample) are available.

The instrument scientist loads the sample changer as specified by the script and starts the instrument control script. The instrument will then collect data on each sample in the sample changer for a specified time and write out the datafiles.

As soon as datafiles are written to disk by the instrument control software/DAE, a file watcher picks these up, extracts metadata and sends this to ICAT. The process then checks if the set of files needed for analysis is complete and if it is invokes an appropriate data analysis/reduction algorithm using the Mantid scripting interface.

When the analysis run in complete, a further Mantid algorithm is started which loads the outputs files into ICAT, creating datasets as appropriate and setting the provenance metadata. Finally this process sends an email to the proposer informing them that their data is available, with a link to the data in ICAT. The data management infrastructure involved is outlined in Figure 19.

The user is encouraged to inform the facility of any publications arising from their use of ISIS. If they do tell the facility these publications are uploaded to the publications catalogue and associated with the proposal

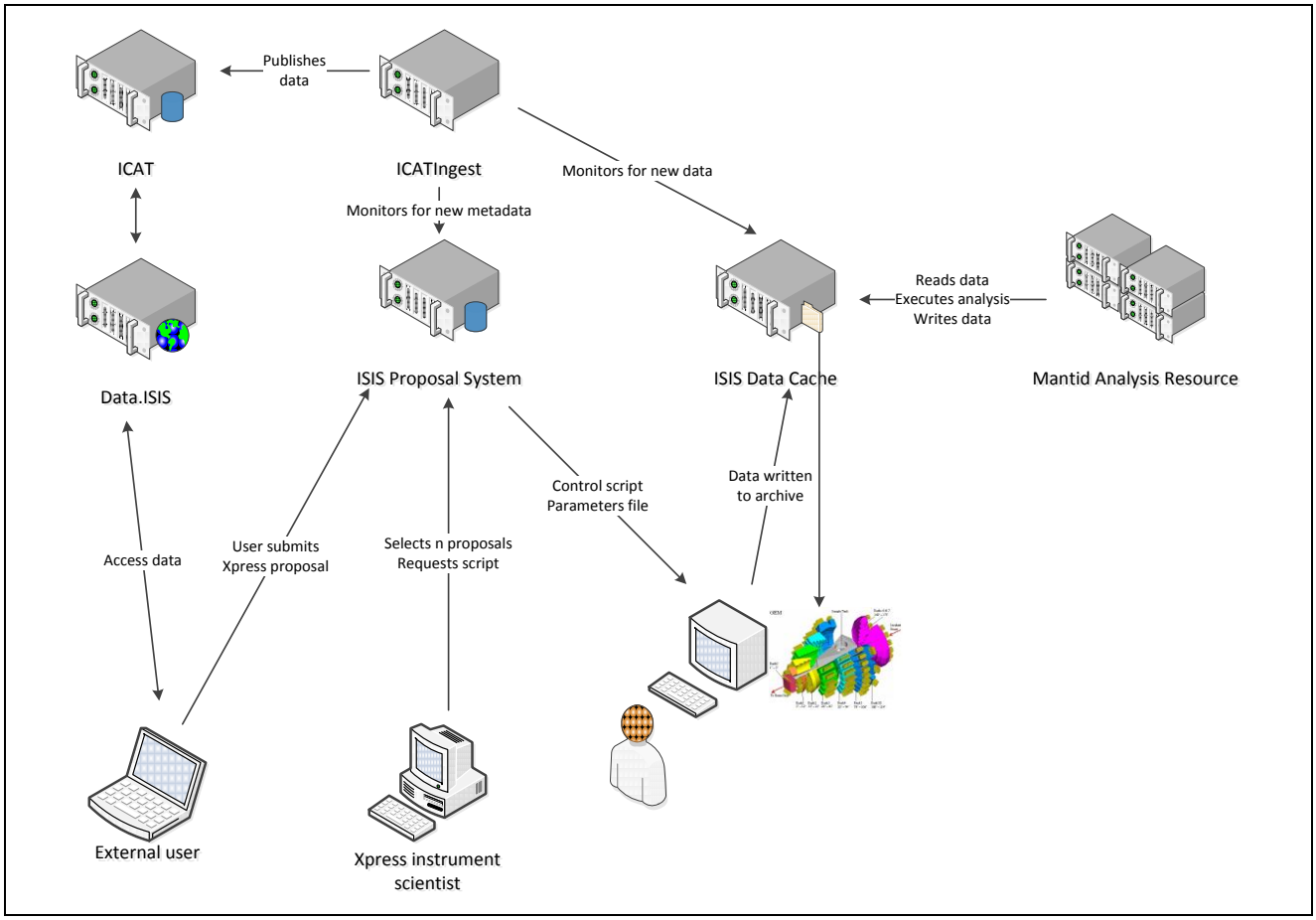


Figure 19: The ISIS data management infrastructure

Stages of lifecycle covered in the scenario:
Proposal submission (sample description metadata), experiment planning, calibration, data collection, data reduction, publication capture. Optionally data visualization.
Data types
• Raw data file (ISIS binary format) are written by the instrument. These are catalogued in ICAT. • Instrument control scripts are a ISIS specific 'open-genie' files • Mantid scripting is controlled by Python files. • Mantid outputs 'gsas' and nexus datafiles, which are catalogued in ICAT
Actors.
• Principal Investigator : Submits proposal. Provides sample. Carries out final structural analysis • Instrument scientist: Reviews proposals. Schedules beam time. Loads sample changer. Initiates data collection • Automated systems: Catalogues raw data. Writes control and analysis scripts. Runs analysis scripts. Catalogues analyzed data.
Metadata requirements
Sufficient metadata to automate the analysis (knowing which files are calibration, empty can, which sample etc) is crucial. The proposal reference number must be known throughout so the data can be made available to and only to the correct users.

8 Scenario 5: Resultant data and publication tracking and linking

Facility: ISIS pulsed neutron and muon source, STFC, UK
Scientific Instrument, technique and discipline: Would cover all instruments, techniques and disciplines supported within the facility.

Stages of lifecycle covered in the scenario:
Proposal submission, experiment reporting, publication recording
Data types.
Proposal information, user information, experiment information, publication information, resultant datasets, supplementary datasets.
Actors.
User office, experiment team, facility library.
Metadata requirements:
Bibliographic record; user identity, experiment identity, resultant data metadata

8.1 Scenario description

As discussed in section 2.3.7 above, one of the steps which the facility require of the scientists allocated time on a facility instrument is to lodge with the facility a record of any publication which was a result of undertaking the experiment at the facility and collecting data.

Publication management is a large area in its own right, and users or library staff will typically refer to the appropriate paper by a bibliographic reference to a paper in an journal archive, bibliographic aggregation service (such as Web of Science), or institutional repository. They may supply the paper's DOI as a short reference to the paper. Such a bibliographic reference then needs to be related to the correct experiment in the facility.

Further, in order to provide an added value service to users and increase the potential impact of the work, and thus that of the experiment at the facility, the facility may also provide a service for the registering and potentially the deposit of the resultant data. By resultant data we mean the final analysed data which underpins the result reported in the paper, e.g. the data behind graphs and tables, or the final molecular structure, often supplied as supplementary data to journal articles.

8.1.1 ISIS ICAT Data Catalogue

ICAT is an open source metadata management system designed for large scientific facilities. It comprises a database with a well defined API that provides an interface to a large facility's holding of experimental data. At ISIS, all experimental data files are produced, captured and catalogued into ICAT along with the metadata about sample conditions for that experimental run, and metadata from the proposal.
The ISIS ICAT implements a web service API which supports different applications for browsing, searching and downloading experimental raw data, e.g. the TopCAT web tool for searching multi-

ple ICAT catalogues (Figure 20, left). ISIS registers DataCite[17] DOIs for experiments which it encourages researchers to cite in publications relating to ISIS experiments. This DOI can be resolved via a handler system to an HTML landing page shown in Figure 20 (right).

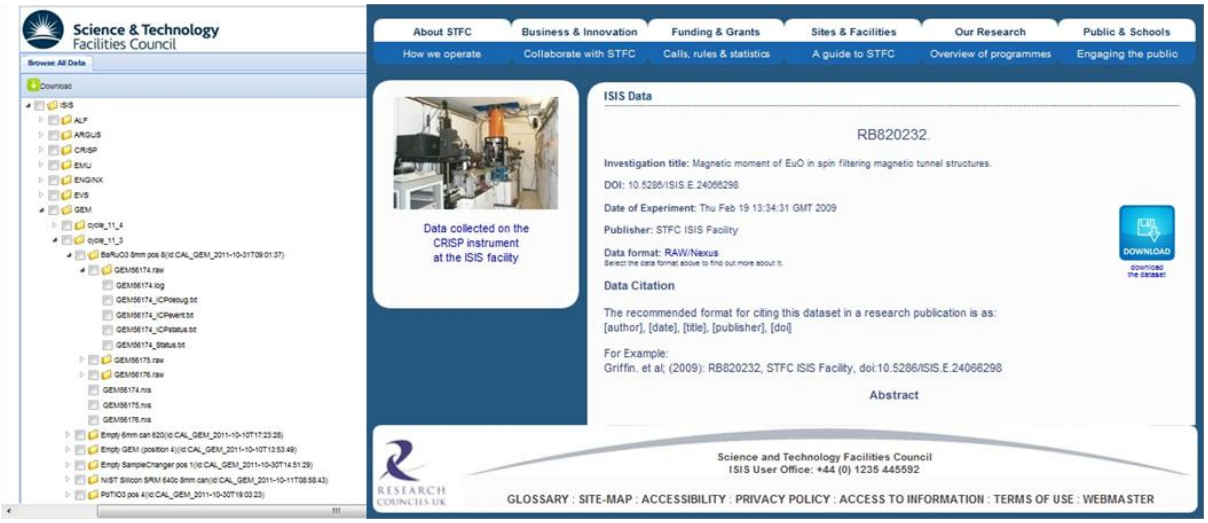


Figure 20: User applications supported by the ICAT API: (right) An ISIS DOI Landing Page, (left) TopCAT web tool for searching multiple ICAT services.

8.1.2 STFC EPublications Archive (ePubs)

EPubs is an institutional repository collecting and providing access to the academic output of STFC, from both STFC authors and also, of particular relevance within this scenario, facility users. It uses an extended version of IFLA Functional Requirements for Bibliographic Records[18] (FRBR) enhanced with Dublin Core elements to permit a wide range of contents to be represented. EPubs content typically include journal articles, conference papers, data, patents, technical reports, ePrints, theses and books. The archive supports open access and is OAI-PMH[19] compliant.

EPubs provides a web interface (Figure 21) for users to browse and search its content. It offers browse indices for author, publication date, organizational structure, material type and full text. Registered users can also use this web application to submit new work and access a range of statistics information, e.g. download count.

8.1.3 Linking Publications and Experiment

Thus facilities such as ISIS wish to form a connection between their experiments and the raw data they generate and the papers, to provide added value and to trace impact. There are several ways in which such a scenario could be described, depending on the deposit route chosen; here we illustrate one possible route[20].

[17] http://www.datacite.org

[18] http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records

[19] http://www.openarhives.org/pmh

[20] Another entry point into this process may be from a search for relevant publications within aggregated publication databases, such as Web of Knowledge (http://wokinfo.com/) or Google Scholar (http://scholar.google.co.uk/) undertaken by Facilities Library or User Office staff. In this case, it may be hard

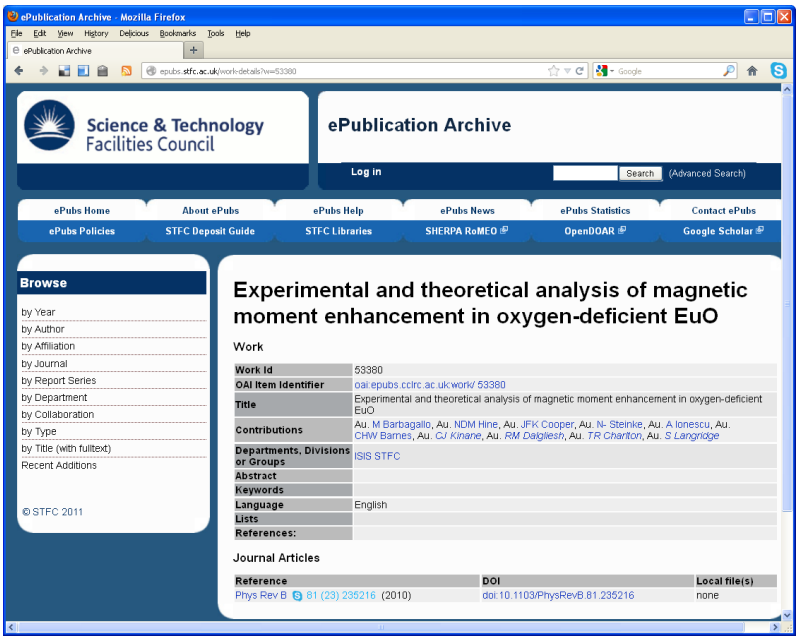


Figure 21: ePubs User Interface

1. The principle investigator is requested to submit publications resulting from experiments. This may either be during an annual request to users for reports on research outputs, or during a subsequent proposal submission when results from previous experiments are requested. The investigator uses a web-based *research output reporting and submission system* to submit: a DOI for the paper; optionally some metadata for the paper; optionally the full text of the paper itself; a reference number for the relevant experiment(s) – either via an internal identifier for the experimental time allocation, or via an allocated DOI.

2. The paper DOI and other optional information is passed to the ePubs system, which dereferences the DOI to access its bibliographic metadata, and creates a record for the paper, linking to the entry within ICAT via the experiment reference number (or landing page of the experiment's DOI); the full text is stored if appropriate.

3. ePubs then passes an identifier to its record to the ICAT API using its publication DOI to create a link from the entry for the experiment to the publication record within ePubs (and its publication DOI). This can be linked from the landing page for the experimental DOI.

4. A report generation system can provide a report on the outputs of the experiment.

to identify which publication refers to which experiment; formally citing a DOI for the experiment and its raw data within the paper would be of great benefit in forming this connection.

8.1.4 Linking to Resultant Data

The facility may wish to also provide a record of the resultant (or supplementary) data. In this case, it will need to provide an additional data storage service to store the resultant data, with its data management and preservation system[21]. In this case:

1. The user would use the *research output reporting and submission system* to additionally either upload directly, or provide a link to a supplementary data set stored elsewhere. The user would be required to provide a set of metadata including: a description of the data, including its format, the (DOI of the) publication it supports and a references to the experiments (there may be more than one) it uses as its source data. References to tools to view and manipulate the data may also be useful.

2. The data would be stored and recorded in the resultant data store and its catalogue[22]. It may be appropriate to allocate a DOI to the resultant data at this stage.

3. A link to the resultant data would be added to the ePubs entry for the publication.

4. A link to the resultant data would be added to the raw data catalogue entry for the publication, with potentially a link from the DOI landing page.

5. A report generation system can provide a report on the outputs of the experiment.

8.2 Discussion

This scenario is different from the other scenarios given in that it does not attempt to capture a complete provenance trail for part of the process, but rather captures research components from different parts of the continuum (proposal, experiment and raw data, resultant data, publication) and forms a link between without necessarily a complete view of the dependencies and derivations of one from the other. The information about the connection between the raw data and resultant data in particular may be partial, as this information may be hard to collect if indeed the scientist is willing to give it in detail. Further, this scenario is less amenable to automation, but would rather require user input, and the tools supplied should be able Nevertheless, there is distinct value to be gained in terms of added value to the end users and the impact assessment to the

This is an area which is being explored in detail within the wider research community. For example the projects: Webtracks[23], OpenAirePlus[24], and Dryad[25] are all making contributions in this area, and we would propose to take advantage of their prior experience in developing a solution.

---

[21] The facility might alternatively rely on safe data storage within external organisations – e.g. university asset management systems. However, such systems are not necessarily in place, may not provide appropriate long-term preservation guarantees and are not likely to be accessible to their own staff. Facilities may also reasonably take the view that it would be more reliable to take an extra copy of the data anyway.

[22] This may be the same catalogue system as the raw data, or a separate one. It may be appropriate to create a separate entry for the resultant data as it could aggregate results from more than one experiment.

[23] http://www.jisc.ac.uk/whatwedo/programmes/mrd/clip/webtracks.aspx

[24] http://www.openaire.eu/

[25] http://datadryad.org/

9 Conclusions and next steps

In this report, we have explored the data continuum for facilities science, with an emphasis on the processes which are undertaken as a experiment is transacted within one facility, from proposal to publication. We have emphasised the data and metadata sources which are available at each stage so that the data continuum can be tracked and recorded, and thus made available for future reuse.

Early stages in the process are relatively speaking within the facility's control, and using the facility's staff and information systems and thus it is relatively straightforward to provide integrated support for those stages of the process. Later stages (analysis and publication) are largely outside the control of the facility, and thus are hard to contain within a single provenance management system. This leads to a careful consideration of the value and costs of managing this information.

Case studies show nevertheless that there are clear cases (and there are further ones which could also be explored) where tracing provenance is of value, and thus generic tools, if they can be developed within reasonable cost, could be developed within PaN-data, which can be used within such scenarios.

In the next stages of the project, we shall look to develop an architecture and information model to support tools which provide support for provenance. These shall include:

- An extended metadata model and ontology which can encapsulate provenance information. Such a model should take into account best practise and standards in provenance and workflow, but be realisable within the practical data management tools within facilities.
- An architecture of tools which support provenance with guidance on how they can be used to support use cases in facilities.
- Proposed controlled vocabulary for common terms used within facilities.

References

[Basham 2012] Mark Basham. "HDF5 Parallel Reading and Tomography Processing at DLS", Joint PNI-HDRI and PaN-data workshop, Hamburg, Germany 2012.

[Crompton et. al. 2012] Shirley Crompton, Brian Matthews, Erica Yang, Cameron Neylon and Simon Coles. "Collaborative Information Management In Scientific Research Process", e-Science 2012, Chicago, USA, 8-12 October 2012.

[Diamond-tomo 2012] Added by Kaz Wanelik, last edited by Kaz Wanelik, http://confluence.diamond.ac.uk/display/BLXIIII/Tomo+Reconstruction+from+Tiff+Files+on+I13. Last access 16 Oct. 2012.

[Glavic and Dittrich 2007] Boris Glavic and Klaus R. Dittrich (2007). *Data Provenance: A Categorization of Existing Approaches*. In Datenbanksysteme in Business, Technologie und Web (BTW 2007), pp. 227 - 241.

[Oinn et. al. 2004] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat and Peter Li, (2004). *Taverna: a tool for the composition and enactment of bioinformatics workflows*, Bioinformatics, Vol. 20(17) 2004, pp3045-3054.

[PaN-data-Europe D7.1]. PaN-data Europe. *Survey of publication repositories, cross-linking and long-term preservation.* Deliverable D7.1, 2011.

[Tucker et. al 2007] MG Tucker, DA Keen, MT Dove, AL Goodwin and Q Hui. RMCProfile: *Reverse Monte Carlo for polycrystalline materials.* Journal of Physics: Condensed Matter 19, art no 335218 (16 pp), 2007, available at: http://wwwisis2.isis.rl.ac.uk/rmc/.

[Yang 2010] Erica Yang. Martin Dove's RMC Workflow Diagram. Project Requirement Report (supplementary report) for the I2S2 project, July 2010. Available at: https://www.jiscmail.ac.uk/cgi-bin/filearea.cgi?LMGT1=I2S2\&f=/Deliverables/RequirementsReport .

[Yang et. al. 2011] Erica Yang, Brian Matthews, Michael Wilson, (2011). *Enhancing the core scientific metadata model to incorporate derived data*. Future Generation Computer Systems, In Press, available online 19 August 2011, ISSN 0167-739X, 10.1016/j.future.2011.08.003.