

Chapter 10

Challenges in Data Intensive Analysis at Scientific Experimental User Facilities

Kerstin Kleese van Dam, Dongsheng Li, Stephen D. Miller, John W. Cobb,
Mark L. Green, and Catherine L. Ruby

1 Introduction

Today's scientific challenges such as routes to a sustainable energy future, materials by design or biological and chemical environmental remediation methods, are complex problems that require the integration of a wide range of complementary expertise to be addressed successfully. Experimental and computational science research methods can hereby offer fundamental insights for their solution. Experimental facilities in particular can contribute through a large variety of investigative methods, which can span length scales from millions of kilometers (radar) to the sub-nucleus (LHC¹). These methods are used to probe structure, properties, and function of objects from single elements to whole communities. Hereby direct imaging techniques are a powerful means to develop an atomistic understanding of scientific issues [1,2]. For example, the identification of mechanisms associated with chemical, material, and biological transformations requires the direct observation of the reactions to build up an understanding of the atom-by-atom structural and chemical changes. Computational science can aid the planning of such experiments, correlate results, explain or predict the phenomena as they would be observed and

¹<http://public.web.cern.ch/public/en/lhc/lhc-en.html>.

K.K. van Dam (✉) • D. Li
Fundamental and Computational Science Department, Pacific Northwest National Laboratory,
Richland, WA, USA
e-mail: Kerstin.KleeseVanDam@pnl.gov; dongsheng.li@pnnl.gov

S.D. Miller • J.W. Cobb
Data Systems Group, Neutron Scattering Science Division, Oak Ridge National Laboratory, Oak
Ridge, TN, USA
e-mail: millersd@ornl.gov; cobbjw@ornl.gov

M.L. Green • C.L. Ruby
Systems Integration Group Tech-X Corporation, Williamsville, NY, USA
e-mail: mlgreen@txcorp.com; rlruby@txcorp.com

thus aid their interpretation. Furthermore computational science can be essential for the investigation of phenomena that are difficult to observe due to their scale, reaction time or extreme conditions. Combining experimental and computational techniques provides scientists with the ability to research structures and processes at various levels of theory, e.g. providing molecular ‘movies’ of complex reactions that show bond breaking and reforming in natural time scales, along with the intermediate states to understand the mechanisms that govern the chemical transformations.

Advances in experimental and computational technologies have lead to an exponential growth in the volumes, variety and complexity of data derived from such methodologies. For example the experimental data rates at Oak Ridge National Laboratory (ORNL) Spallation Neutron Source (SNS) vary from around 200 MB/day to around 4.7 GB/day per instrument with an average of around 1.3 GB/day/instrument for its 23 instruments. The Advanced Photon Source (APS) has almost 60 beamlines with one to three instruments per beamline and rapidly produces copious amounts of data. Typical experiments will produce between a few KB to 100 GB, while imaging experiments (such as tomography) produce more, on the order of 1–10 TB of data per experiment. Data rates for some instruments, such as X-ray Photon Correlation Spectroscopy and 3D X-ray Diffraction Microscopy will approach 300 MB/s on a continuous basis. At the Linac Coherent Light Source (LCLS) there will be six sets of versatile high data bandwidth instruments installed in two hatches of the LCLS experimental area. Some instruments will be capable of producing up to tens of GB/s of data in peak. In the final implementation of the system up to two of those instruments can be used simultaneously. This will result in multi-terabyte data volumes to be handled on daily basis. The data rate will ramp up over the next several years. The first experiments will produce up to 1 TB of data per day. In 3 years the amount of data to be stored per day will increase up to 15 TB from only one of the instruments, and that would correspond to nearly 2–3 PB of data per year. Next generation facilities such as the X-Ray Free Electron Laser in Germany (XFEL²) expect data rates of up to 3.5 PB a day, compressed and reduced for long time storage to 1–4 PB a month [3], in comparison the much quoted LHC particle physics experiment is expecting to store 5 PB a year. However it is not just the large scale facilities that have experienced this increase in data rates, facilities with laboratory based equipment such as the Environmental Molecular Sciences Laboratory (EMSL) with over a hundred different instruments have seen similar increases. A 2010 Ion Mobility Spectroscopy Time of Flight instruments produces 10x as much data as comparable systems in 2006 i.e., an increase from 1 to 10 TB per day.

Similarly submission rates at leading community repositories for experimental data results such as the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL-Bank – comparable to the US GenBank) have strongly

²www.xfel.eu.

increased. EMBL is currently growing at a rate of 200% per annum, requiring a doubling in its storage capacity every year (5 PB in 2009). EMBL's web pages, which give access to 63 distinct databases, received ~ 3.5 million hits per day in 2009, and programmatic access to their data via web services was at ~ 1 million requests per month and growing [4]. The number of databases reported in *Nucleic Acids Research* jumped from 218 to 1,170 in the years 2001–2009 [5]. Overall experimental data rates have significantly increased in line with the general trend for new experimental instruments; as a consequence, whilst the data deluge might not be happening everywhere in an absolute sense, it is in a relative one for most research groups. However it is not only the volume that has increased it is also the complexity of the data that is rapidly growing. Ever new investigative methods are being developed, with each method and vendor of instruments for such a method creating new data formats and results representations. Therefore experimental science and its analysis is overall now a very data intensive field of science.

This exceptional growth in data volumes and complexity has presented researchers with significant challenges, foremost how to effectively analyze the results of their research both for single experiments and increasingly across different investigative methods. The availability of underpinning data management facilities and tools play hereby a crucial role throughout the experimental and analysis processes.

Data management challenges include issues such as data storage, access, and movement. Ever growing volumes no longer allow facilities or researchers to store all the collected raw and derived data in perpetuity and hard decisions might have to be taken in terms of what is worthwhile retaining. Even when it is possible to store the data collected, its volume and diversity requires expert management to enable immediate and timely analysis as well as long term access and usability of the data, this data management knowledge is not always available at the researcher or even the facilities level, leaving large volumes of data destitute and inaccessible. Similarly it can be quite difficult for scientists or facilities to support the basic functions necessary for the correlation of research results. Data transfers between organizations can be fraught with problems such as unreliability, speed and lack of data integrity throughout the transfer, and so many facilities and their users still rely on the shipping of hard drives for their data movement [6].

An even greater challenge however is the analysis of the data itself, with the increasing variety of instruments used at experimental facilities; the variety of (often proprietary) data formats and analysis software packages has increased dramatically. This plethora of investigative methods and data formats has prevented the community thus far from working collaboratively on advancing their analytical methods. As a result traditional analysis methods are often not scalable enough to deal with either the increasing volume or complexity of the results of both experimental and computational research results. In response researchers often either do not use the full capabilities of the instruments or only analyze a very small subset of the data they collected. Where full analysis is possible it can take

many hours or weeks for an experiment which might have only lasted minutes or seconds. The lack of suitable tools, advanced computing techniques, and storage are key limiting factors. Furthermore these hinder the progression of the field to address one of the main requirements for the future of experimental science: the ability to analyze experimental results in real time, and actively influence these. To achieve this next level of experimental research, a new generation of analysis methods needs to be developed.

Experimental science today is highly specialized on the individual level, driven by ever more complex investigative methods, but very collaborative and international in its project work [7], driven by the complexity of the scientific challenges. It is therefore necessary to correlate, integrate, and synthesize local results with other experimental and computational work world wide to improve the quality, accuracy, and completeness of the analysis. Therefore a critical challenge for experimental science is the need to empower the scientist with computational tools to perform analysis across a high volume, diverse and complex set of experimental and simulation data, to extract the desired knowledge and meaning that leads to scientific discovery.

This chapter will discuss the critical data intensive analysis challenges faced by the experimental science community at large scale and laboratory based facilities. The chapter will highlight current solutions and lay out perspectives for the future, such as methods to achieve real time analysis capabilities and the challenges and opportunities of data integration across experimental scales, levels of theory, and varying techniques.

2 Challenges

The experimental sciences community faces a wide range of challenges, both in their day to day work, as well as in their endeavor to progress the scientific capabilities of this domain in general. While many challenges are related to the instrumentation, the specific science domains or physical research objects, an increasing number stem from the data intensive nature of the processes involved in experimental analysis. In the following we will elaborate further on the key challenges which include the following principle areas:

- Metadata Generation and Association
- Data Formats
- Data Integrity
- Data Analysis of Single Experiments
- Co-analysis of the Collection
- Data Provenance, Collaboration, and Data Sharing
- Data Ownership and Data Citation

2.1 Metadata Generation and Association

Often experiments require the use of specialized sample environment equipment to control the conditions in which data are collected. The nature of the sample environment equipment may vary widely though the most common equipment controls temperature, pressure, or magnetic field, and sometimes a combination of these. Other common types of sample environment equipment may include pulsed lasers, vibrational stress to materials, dynamically mixing gases or chemicals at varying ratios, and the sequencing of a number of samples to be studied. The timing for when conditions change on the sample must be recorded in order to correlate sample environment changes with those observed in the data, and are critical to the reliable analysis of the results. The how precise the correlation needs to be can depend upon the time scale for the rate of change anticipated. This being the case, it is possible that the sample environment metadata can also be an appreciable size.

Similarly, sample positioning is another important piece of metadata which must be recorded. The position information is necessary to have for experiments which rely upon probing beam path lengths and angular relationships to detectors. The position information is also necessary for classes of samples which have positional information as part of their composition, such as crystalline materials. Often these crystals must be oriented according to a known position so that structure can be studied. Another class of experiments fall into the category of rotating the sample such as in the case of tomography or neutron spectroscopy inelastic single crystal energy transfer measurements. Moving or rotating the sample can create a corresponding data file for each positional change, else the dynamic nature of the repositioning must be indicated within the data.

There are a wide variety of other metadata which should also be associated with the experiment and which metadata are recorded can be a function of facility capabilities and data policies. Often one primary key is the experiment proposal number. Along with this, associating the experiment team members can be important. Other important metadata include: measurement start and end times, instrument status and operating conditions, and data acquisition related metadata used to properly identify and segment data such as measurement frame numbers or experiment run numbers.

Many of these metadata are not only vital for the immediate analysis process, but also support the long term exploitation of the results. The quality and comprehensiveness of the metadata will directly influence the accuracy and quality of the analysis process. Due to the wide variety of metadata and its sources its structured and quality controlled capture is a major challenge for any scientist or facility.

2.2 Data Formats

Data can appear in a wide variety of formats ranging from unformatted, proprietary, to self-describing such as HDF5³ or NetCDF⁴. Data management professionals can be challenged to select the best data format to use for a number of reasons ranging from ease of use to file format performance. Scientific communities can engrain on particular data formats, which can cause challenges when data systems professionals seek to use a new format. There can be little motivation to adapt community developed legacy applications to utilize new file formats, particularly if these new formats are not readily available via the software language used by the legacy application. Scientists may be familiar with their own tools for examining data and if new formats cause any additional burden, these will find low acceptance.

However one should not give up on defining file formats, particularly in scientific communities lacking data format standards. Moving the community towards standards has the benefit of potentially opening up software development to a broader segment of the community once data interchange formats have been established. In some cases, it may be necessary to move the community to more advanced data formats to address issues that might have already been solved by these standard data formats. For example, if higher performance is needed, utilizing the parallel access or inherent data compression/decompression capabilities may be of benefit.

The longer term benefit of using a self-describing data formats are many, as the file can collect and store metadata pertaining to the data which may otherwise be lost over time if these are maintained as separate files. A self-describing format also offers the potential to engage a larger community of researchers wishing to collaborate on the data. Thus there are many advantages for defining data formats for a scientific community.

Another challenge for establishing data formats may be to capture data with non-traditional data formats. Typical data acquisition may utilize histogramming to bin data. However, the binning process can reduce the resolution of the data. To avoid this problem, some state of the art instruments are utilizing *event mode* data which works by concatenating detected data to a growing list. One such example would be an event data format which records detected events in position and time thus providing the maximum possible resolution of the detection system. However some data storage formats may not respond well to varying length data sets, especially if data need to be appended during an acquisition.

³The HDF Group produces and maintains software for self-describing scientific data via the Hierarchical Data Format. <http://www.hdfgroup.org/>.

⁴The Network Common Data Form (NetCDF) self-describing data format developed by the University Corporation for Atmospheric Research (UCAR). <http://en.wikipedia.org/wiki/Netcdf>.

2.3 *Data Integrity*

Ensuring the quality of the data is a challenge as part of the data acquisition process for one primary reason – ensuring data integrity may take as much time as the time required to produce the data, and the data acquisition system may not be capable of performing this task during rapid data acquisition sequences. However the challenge remains that data integrity must be ensured as close to the source as possible.

There are a number of data integrity mechanisms that can be employed, some implicit while others are explicit. Implicit mechanisms include computer memory parity checking and correction, network handshaking protocols such as TCP/IP, and data parity checking and correcting methods such as disk systems which utilize RAID parity checking and correction, such as RAID 1, 5, 6, or 10. These methods are commonly utilized today and are fairly reliable and robust enough that one may take for granted that additional data integrity mechanisms may need to be employed.

However if one does not explicitly determine the integrity of the data, one may not know for sure the integrity of the data. Considering the vast sizes of data sets today, the probability for some type of data corruption is on the increase. These errors may arise from faulty memory, RAID system failures or single disks not operating in RAID configuration, or faulty networking equipment that corrupts data during transfer. In the case of an unnoticed error resulting in data corruption, the corrupted data may be perpetuated into the future beyond a point where it can be recovered.

To help explicitly identify data, methods for producing checksums have been developed. A checksum is typically a fixed-size number computed from a data set of interest that, to some high degree of certainty, uniquely identifies that particular data set. Thus a change in one of the datum will result in a different checksum. A variety of checksum methods are in use with the more common ones being MD5⁵, SHA⁶, and CRC⁷.

When examining the dataflow, ideally the checksum process must be performed as early in the dataflow as possible and as previously mentioned ideally when the data are created. To be useful, the checksum must be stored somewhere where it can be referred to at a later time. Data production systems often employ catalogs to store metadata for search, and the checksum value for the data should be stored in this catalog. Though the challenge remains today, to create checksums in a timely fashion for large files.

⁵<http://en.wikipedia.org/wiki/MD5>.

⁶http://en.wikipedia.org/wiki/Secure_Hash_Algorithm.

⁷http://en.wikipedia.org/wiki/Mathematics_of_CRC.

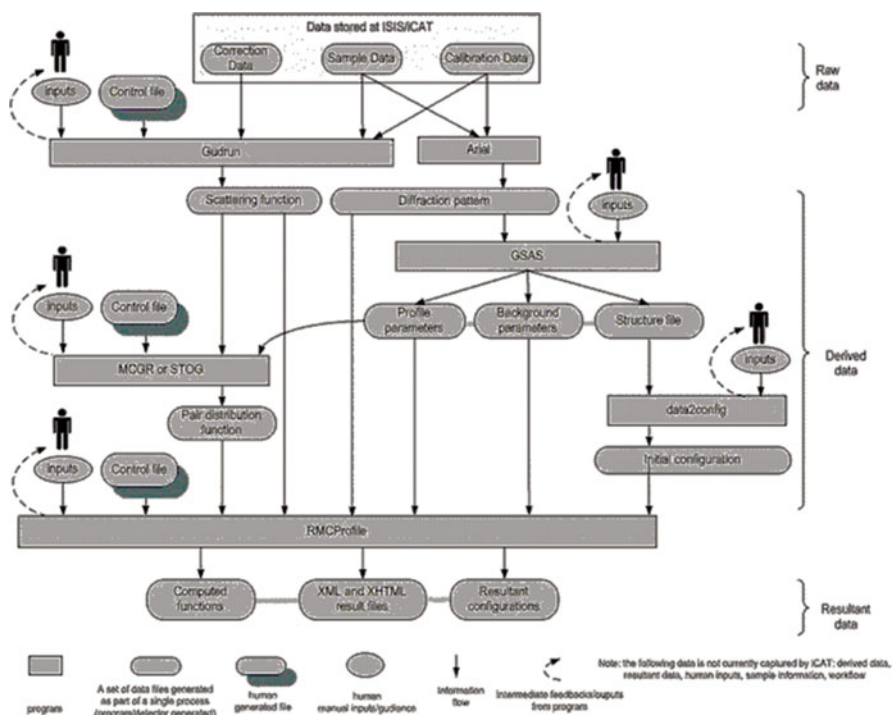


Fig. 10.1 Example crystallography analysis workflow [8]

2.4 Data Analysis of Single Experiments

The analysis of the raw data produced by single experiment is often a complex process, incorporating many different steps, some of which will need to be repeated several times over after review, to achieve the best possible results (see Fig. 10.1 for an exemplary analysis workflow).

The steps taken can in general be classified as: data capture, calibration, data compression, data reduction (Reduce noise and smooth data, reconstructions will contain the most significant information, are feature-accentuated), image reconstruction (Accurate re-construction of high volume data, combine correlation functions with parallelized filtered back projection), segmentation and feature association (identification of application-specific chemical signature and feature recognition), visualization of results. Some of the analysis steps might be repeated several times to identify all required features in the data and filter out enough background information to make these clearly visible.

The increase in data and repetition rates on many instruments has caused severe problems for the subsequent analysis. The analysis take much longer than before e.g., up to 18 h for a basic analysis for a mass spectrometry experiment that takes

itself under 1 h. But more importantly many of the existing tools are no longer able to cope with the data volumes, requiring the scientists to collect data with less precision than the instrument could offer or being able to examine only on small subsets of the sample (i.e., $15 \times 1,000$ rather than $1,000 \times 1,000$) thus hampering their scientific research significantly. The problem in the existing methods are hereby not only the data throughput, but the mathematical methods used, many of which do not scale or are not at the appropriate level of theory. An example repetition rates at leading, large laser facilities have increased from one shot an hour to one a second, whereas in the past direct analysis methods where appropriate scientists now need to investigate much more complex methods, separating the effects of different shots, but also consider more statistics based approaches to their analysis. Later stages of the analysis such as segmentation and feature detection face similar challenges due to the increased volumes and complexity of the results. Current more interactive methods of feature identification need to be replaced by automated ones. More importantly the representation of results has become more challenging; high levels of details in single visualization (e.g., 3D volume rendering of dense bio films) make it difficult for users to locate features of interest. The data volumes are so big that only very advanced visualization tools can cope, however, these are often difficult to handle and require specialist support, traditionally not present at experimental facilities or the scientists home organization. Similarly if the users want to interact with the visualization, the data volumes require significant processing power to support this interaction. This processing power can no longer be provided in the traditional analysis setting at the researcher's desktop, but requires dedicated visualization clusters and specialist software e.g. for remote visualization back to the researchers desktop. While such methods exist, these tools are made for visualization specialist in main and not for the use by scientific end users.

Driven by the need for science-enabling tools and technologies, researcher are increasingly interested in real-time analysis and visualization e.g. of protein crystal and enzyme structures and functions to enable intelligent instrument and experiment control. It has proven particularly successful to pair researchers with computer and computational scientists. These can guide researchers through structured requirement gathering exercises, to identify enabling technologies that would address their needs and provide a real step change in the possible scientific analysis. In the recent commissioning of the neutron science instruments at ORNL's SNS the computing team heard questions like the following that could be answered by the right computational infrastructure:

- If I could plot in real time R-factors and observations-to-parameters ratios, they should asymptotically approach values limited by the sample. I could then see when the best time is to stop the data collection and move to the next temperature or the next sample.
- Say I want to know the O – H, O hydrogen bond distances with a precision of 0.01 Å. If I could evaluate bond distances their esd's in real time, I could see if and when this is achievable.

- Parametric studies with single crystals – observing the dependence of a structural parameter versus time, temperature, pressure, magnetic field, electric potential, laser excitation, and gas diffusion.
- Observe Fourier density maps in real time.
- Follow an order parameter in an order-disorder phase transition in real time.
- Follow the intensities of superlattice and satellite peaks and diffuse scattering in real time (reciprocal space).

However, the understanding of how to use leadership computing facilities as part of such a computational infrastructure can be extremely time consuming for the scientists to learn. Moreover, the access to these world-class resources is highly dependent on the physical location, network connectivity, local computer resources, or other resource-limited device availability. Leading community support facilities now provide scalable user access through thin- and thick-client computing models. Thin-client access is generally suitable for handheld resource-limited devices and/or local computer resources where it is advantageous for the application software, data, and CPU power to reside on the network server rather than the local client. In most cases this will require that the scientist has network access and can access a web browser, and no further application software installation or support is required. Conversely, thick-client access is highly desirable when application software, data, and CPU power is provided by the local resources that are able to function independently from the network server. Portals and Science Gateways can supply a robust support infrastructure for clients of this type by providing resource and application availability dependent on the user requirements and level of sophistication.

There exist several neutron instrument simulation packages such as MCSTAS, IDEAS, NISP, and VITESS, which are used by instrument scientists to develop detailed simulations of their beamlines. While in some cases several man-years of effort are invested into these simulations, these valuable models are not routinely being used by the neutron scientists for virtual experiment optimization and planning due to computational and data workflow complexities. Furthermore, a current bottleneck of efficient data collection is the lack of software allowing for real-time tracking of diffraction patterns as they are being collected in an integrated manner. Current single crystal diffraction instrumentation designs will be able to survey a vast amount of reciprocal space within a short time period. The data produced, composed of short Bragg reflections and diffuse scattering, carry important information on static and dynamic interactions of molecules and atoms in the crystalline state. Computer programs such as GSAS, Fullprof, and SHELXL are readily available for post-mortem data analysis interpreting Bragg diffraction data and refining ordered structures. However, a real-time system would enable biomedical structure refinement to occur while samples are still in the instrument. This would provide a real-time determination of experiment duration based on refinement criteria provided by the scientist and verified by the real-time analysis of live experimental data, which has never before been attainable. Enabling a real-time decision support system for neutron beam experiment users has the potential to dramatically advance the state-of-the-art and lead to not only the more efficient

use of facility resources but may also lead to a better understanding to the dynamics of data collection within an instrument.

2.5 Co-analysis of the Data Collection

As in the medical field where a doctor may order a number of tests and scans to examine and diagnose a patient, so do scientists and researchers utilize a number of experimental techniques to study objects. In the case of large experimental user facilities, it is quite common for scientists to perform complementary experiments at both X-ray and neutron scattering user facilities, or combine laboratory based experiments with X-ray experiments. Often the X-ray data can give good information regarding the structure of a material, while the neutron scattering data can give valuable information on the placement of hydrogen atoms within this structure that X-rays see quite poorly. Laboratory based instruments might give more information on the chemical composition of the object or its functions. In other imaging technique combinations one technique might provide a cost effective first quick look at an object, whereas another one is used to examine objects of interest identified in the initial quick look in much more detail with a higher precision method. Therefore these techniques can complement each other quite well in providing different pieces to the puzzle much like the various tests that a medical doctor would have performed for a patient. Furthermore the results of one experiment might not only inform the planning of another experiment, but can help in its direction and analysis.

While a few software programs are emerging for specific imaging technique combinations e.g., to co-analyze some X-Ray and Neutron experiments, most of the co-analysis is currently carried out in an *ad-hoc* fashion by the researchers and thus is very time consuming and error prone. The challenges in this type of co-analysis do not only lie in the analysis algorithms, but equally in the vital logistics support for this type of analysis, with data often residing in different institutes and potentially owned by different users.

The required data management of such complementary data sets are typically left *completely* to the user. Often these experimentalists must manage copying data to portable disk drives while they continue to acquire more experiment data. The portable disks tend to be low performing and may be the only copy of the data resulting from the experiment as facilities typically have a short time on the order of two weeks when they will keep data available for the facility users. Some facilities have developed more mature data management practices and systems, and retain data for longer periods of time to better facilitate the data reduction and analyses processes inherent in the publications process.

Thus the experimentalist must deal with a variety of factors including:

- Different or no data management systems at one or both user facilities
- Different computing access mechanisms for each facility perhaps resulting in multiple passwords to manage
- Managing where data reside

- Single copies of data are vulnerable to errors and loss
- Resource limitations for slow performing data systems and computers

Once the results of all experiments are available, the real analysis challenges start, as there are hundreds of different investigative methods, the user has to determine how the different imaging techniques relate to each other and thus how they need to be treated. Do the results need to be integrated, compared, correlated etc. The representation of the results from different techniques varies significantly, as does their scale, accuracy, and measured property. To compare two experimental results, experts in both techniques need to be present to determine the relationship and necessary analysis steps for their co-analysis, giving the lack of available tools, they would then need to develop the algorithms to carry out the analysis and evaluation of the results. Increasing data volumes and experimental complexity have made this type of co-analysis ever more challenging and thus deterring many. Where scientists embark on this journey, it will take them many weeks or months to complete. Given that their foremost interest is the scientific outcome, the tools produced are *ad-hoc* solutions, usually fit only for this specific analysis and not ready to be shared with others. More importantly, most of the time they will have no means or interest to share their methods, thus other researchers will have to start again from scratch, would they decide to follow in a similar direction.

Thus it is evident that there are many barriers to multi-facility data collection and analyses. However, the rewards for improving inter-facility data management and co-analysis software tools may likely yield an accelerated pace for scientific discovery for many science areas. Though this has been an area which has been slow to advance due to the complexity of coordination required for inter-facility data management, and in some cases, it is more a matter of policy than technology which may impede this integration.

2.6 Data Provenance, Collaboration, and Data Sharing

Research projects are increasingly looking for ways to effectively keep abreast with everyone's progress, discuss findings, share data, applications and workflows and manage the documents that are exchanged, before publishing their results to a wider community. Tools like dropbox, Google groups, Google docs, megaloader, etc., do allow exchange of data, but they fall short in the following areas:

- Limited space with paying a subscription fee. This becomes difficult when team members all need to subscribe.
- These tools do not provide the version control and tracking capabilities that are needed.
- These tools do not allow you to annotate the data files and attach discussion threads to datasets.

Wiki's, another popular choice, become unwieldy very quickly and are not suitable for large data exchange. Existing repositories at experimental or computational facilities or community archives, provide access to data, but offer no support for wider reaching scholarly exchanges and collaboration. There are only a few notable exceptions where this kind of project based management, exchange and sharing of data is supported by a community wide infrastructure, these are: the long standing Earth Systems Grid (ESG⁸) offering data exchange and sharing for climate simulation data worldwide, the relatively new NFS funded I-Plant⁹ Infrastructure's project and community based collaborative data sharing environment and the planned DOE Systems Biology Knowledgebase.¹⁰ Therefore at present most research groups have to rely on *ad-hoc* solutions, which are often difficult to maintain and not efficient.

A further challenge in collaborating both within research projects and across projects is the lack of ability to transfer the increasing amounts of data produced at experimental facilities. Due to the complexity of the network infrastructure, ad hoc nature of the transfers, data sizes and the interaction of end user applications and networking are all contributing factors to this situation. The severity of the networking challenges faced by the users varies depending on the size and rate of the data to be transferred and the regularity of the transfer. Small scale transfers (MBs to a few GBs) are relatively well supported today, although data collection in the 'field' is still a challenge. Medium range transfers (few tens of GBs) can be unreliable (e.g. lost connection), even more so when these are used for data streaming (sequence of experimental or observational measurements). For large-scale data transfers it can be very difficult and time consuming to resolve network problems between any two sites. There are usually multiple carriers that are participating in the end-to-end network path, and it is difficult to get any one carrier to take ownership of the problem. Experiences have shown that to "clean-up" a connection can take in the worst case several months. So if a connection is not of useful quality, it is usually going to take days if not weeks to resolve the problem. In this case the researcher would probably either find a work-around (i.e. send it in the post) long before the problem was resolved or give up, if this was an ad hoc requirement [6]. Therefore new means would be required to co-analyze the results, without the need to move the data.

When data finally ends up in a publication perhaps as a chart, graph, or image, the researcher needs to feel a high degree of confidence in being able to reproduce the results. To do so, the researcher not only needs to be able to refer back from the publication data to the analysis data, to the reduced data, and finally to the acquired data, but also to the processes used to derive the different results – thus the researcher needs access to the provenance of any published data. This is quite a complex chain once one takes into consideration that a large number of separate data products that may have been used in conjunction with the experiment dataflow. Facilities can help

⁸www.earthsystemgrid.org.

⁹<http://www.iplantcollaborative.org/>.

¹⁰<http://genomicscience.energy.gov/compbio/#page=news>.

with the cataloging of acquired data and accompanying provenance information, and possibly with the cataloging of reduced data if this data was reduced using facility resources on-site. However data analysis is typically on the leading edge of scientific discovery and often this is where scientist and researchers utilize a wide variety of tools including software they produce for themselves which is almost impossible to keep track of.

2.7 Data Ownership and Data Citation

So who owns the data? This is a commonly asked, and sometimes hotly debated question. In the case of national user facilities, the government funded the operation of the facilities and one may think that this makes a clear case for data ownership. However oftentimes the people performing the experiments also apply significant skill, labor, and expertise to produce the sample they place in the beam in order to produce their data. Thus making the data openly available immediately could be a significant demotivating factor perhaps fostering a counter-culture of parasitic research.

The case of data ownership and access typically needs to be established by the facility via the data practices and policies which they assert. A one-size-fits-all policy across user facilities may not be appropriate as there may be different factors to consider such as the reproducibility of the experiment and the data, the data volumes produced, the longevity of usefulness of the data, and the typical publication lifecycle for a particular technique – there are many more considerations here. However it is typically agreed upon that at some point experimental data should become publicly available after some predetermined amount of time, though this means for opening data to the public is not universally applied.

Should these data become public, there typically are no standards pertaining to how to cite this data. One means has been to keep the data closed and perhaps only provide it to collaborators who agreed to include the experiment team or the Principal Investigator on the resulting paper. This method has its merits for ensuring data citation, however working this way could also impede the scientific discovery process by not allowing broader access to the data. Data ownership and data citation become most contentious when “hot topics” in science emerge. For example, in the current situation of working to find high temperature superconductors, competing research teams do not want to give away their advantages – or their data.

Stepping back and surveying the scientific data management community, there are emerging standards called a Digital Object Identifier¹¹ (DOIs) which is a character string that is used to identify a data product. The DOI and its metadata may also include the data URL where a researcher may locate the data. The DOI

¹¹http://en.wikipedia.org/wiki/Digital_object_identifier.

system is implemented via federated agencies coordinated by the International DOI Foundation¹². Some thought needs to be given for how to define DOIs for data products as one may easily define the DOIs either as too fine grained or too coarse grained. Other complications are what to do with DOIs in the case where data are adapted from the original – should a new DOI be defined or should the original DOI stand? The answer depends upon the context of the data situation. However there is a well-established method via DOIs that could be employed to help with data citation, though it is far from being universally adopted amongst user facilities.

2.8 Summary

Data intensive analysis at experimental facilities faces a wide array of challenges, chief amongst them:

- Current algorithms are often unable to handle the increasing volumes and diversity of the data either at all or in a timely fashion.
- The community requirement for real time analysis cannot be met with present solutions.

In addition experimental analysis relies heavily on the integration, correlation, comparison and synthesize of the single experimental results with other experimental and computational efforts, requiring not only multi-modal analysis and visualization solutions that can span scales, levels of theory, and investigative methods, but also a supporting eco-system of data management, movement, security, and collaboration services to enable this type of co-analysis.

3 Current Solutions and Standardization Efforts

Many of the challenges described in the previous section have been known to the community for a considerable time; however the pressure to address them has only increased in recent years due to the exponential increase in data volumes and the drive for co-analysis of results. Community efforts so far have largely concentrated on the improvement of data management support at experimental facilities and the optimization of single experiment analysis. A few small developments are emerging at present in the field of collaboration support for experimental sciences. In this section we will describe some key developments in these areas, exemplary for the field.

¹²<http://www.doi.org/index.html>.

3.1 Data Management Practices and Policies

As data are a fundamental product of the user facilities, by default *de facto* data management practices will evolve, but in more deliberate and formalized situations, data policies are defined and put into practice. In surveying a number of DOE user facilities, it quickly became apparent that as of this writing the data practices and policies vary widely. Generalizing across the big data producing facilities, the newer facilities appear to be taking on some form of data management for their facility users while the more established facilities (over 10 years in operation), tend to provide less data management resources. It is important to keep in mind that data storage capacity and network bandwidth has increased dramatically over the past 10 years, and this increased value per unit capacity allows facilities to consider providing more services to users, with the goal being to accelerate the rate of user produced publications via data management and data analysis services. To this end, some of the newer facilities have created data centers for their storage needs. The Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory¹³ has a 2 PB parallel file system in their instrument hall [9]. Similarly, the NSLS-II data facility once fully built estimates that aggregating across its 58 beamlines that the facility could produce up to 500 TB per day which via the technologies available today would be completely impractical to utilize data practices based upon portable media for dissemination of experimental data. The Spallation Neutron Source¹⁴ at Oak Ridge National Laboratory has had a functioning data portal coupled with computing infrastructure since 2006 which utilizes a data management system layered upon centralized data storage [10].

Also important to consider are country specific guidelines and policies such as the US Federal guidelines and standards on information management as put forth in FIPS Publication 199 [11]. The security objectives which are of concern:

- Confidentiality – “Preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information. . .”
- Integrity – “Guarding against improper information modification or destruction, and includes ensuring information non-repudiation and authenticity. . .”
- Availability – “Ensuring timely and reliable access to and use of information. . .”

The impact of a breach of confidentiality, integrity, or availability is assessed to be either: low, moderate, or high, depending upon the level of adverse affect on the organizations operations, assets, or individuals. Typically in the case of open research, the impact is assessed as low impact.

¹³SLAC: <http://slac.stanford.edu/>.

¹⁴SNS: <http://neutrons.ornl.gov/facilities/SNS/>.

Standards to define harmonized data policies across user facilities are forming in Europe via the PaN-data a Photon and Neutron Data Infrastructure collaboration.¹⁵ Currently there are 11 partners in the PaN-data collaboration from across Europe. The PaN-data collaboration strives to produce a sustainable data infrastructure for European Neutron and Photon laboratories with goals to produce a common data infrastructure across the large European user facilities that supports the scientific communities in utilizing these facilities. The work being done by PaN-data includes standardization activities in the areas of: data policy, information exchange, data formats, interoperability of data analysis software, and science lifecycle integration of publications and data.

The PaN-data policy document,¹⁶ under development for approximately 18 months, was finalized in December of 2010. The document standardized upon NeXus/HDF5 for data formats. The document also strives to strike a balance between the competitive and collaborative nature of scientific discovery. The open access data policy is intended to provide raw data for use and scrutiny by other researchers, enable data re-use without the need (and additional cost) for re-measuring, and facilitate data mining to facilitate new research.

Examining key PaN-data policy elements:

- Data produced at publicly funded facilities are open access with the facility acting as the custodian.
- Data are to be curated in well-defined formats.
- Automatically captured metadata shall be stored and utilized to form a data catalog which will be on-line searchable.
- Data are provided as read-only.
- Ideally each data set will have a unique identifier.
- Access to raw data and associated metadata becomes open access after 3 years from the end of the experiment.
- Appropriate facility staff will have access to the data.
- The Principal Investigator has the right to copy and distribute the raw data and can grant access to others.
- Ownership of results from the analysis of the raw data depends upon the contractual obligations of the researchers who performed the analysis.
- The facility will provide the ability for users to upload associated results and metadata.
- The facility cannot be made liable in the event of data loss or unavailability.
- Publications related to experiments performed at these facilities are to be made known to the facility within 3 months of the publication date.

In the case of proprietary data where the user does not wish for the data to be made publicly available, beam time is to be purchased at a rate to be determined by the facility. One could expect such fees to be on the order of some number of thousands

¹⁵PaN-data: http://www.pan-data.eu/PaN-data_Europe.

¹⁶PaN-data Data Policy: <http://www.pan-data.eu/imagesGHD/0/08/PaN-data-D2--1.pdf>.

of dollars per hour keeping in mind that an experiment typically lasts from 1 to 3 days.

To support operations, US user facilities either formally or informally have developed data management practices and policies. Typically the biggest difference from the PaN-data policy has been in the areas of data ownership and access, as the raw data are not obliged to become openly available. However advancements are being made in the area of Scientific Data Management (SDM) as an inter-agency working group has been producing recommendations and guidelines [12]. Outcomes from this working group include:

- Agencies should stimulate cultural change through a system of incentives to stakeholders. SDM policy should motivate agency researchers to move from the ownership mindset of data hoarding to a data sharing approach.
- Each agency should develop a data policy within a federal policy context.
- Agencies should manage scientific data for appropriate control while ensuring appropriate access.
- Agencies should establish the roles of chief data officer and should clarify roles and responsibilities.

3.2 Data Management Infrastructures

Experimental facilities support a significant stretch of the experimental research process (see Fig. 10.2).

After a successful proposal for experimental time, the researcher will work with the facility on the experimental design, including instrument configuration and mode of experimental work. For more standardized measurements such as crystallography or proteomics, samples are usually sent to the facility, experimental data is collected, raw data analyzed, and processed data is returned to the user. The majority of experimental work however requires the presence of the scientists at the facility, working hand in hand with the local instrument expert on the experimental set-up, data taking and analysis. Key to the effective support of these processes is the easy availability of information, tools, and data that are required for each step. This required information can include not only data and metadata generated at the facility itself, but also other resources such as data from previous experiments at other facilities, new tools or discussions about experimental set up. The increasing complexity of the processes, and a drive to higher efficiency in the facilities operation lead in the early part of this century to the development of concepts for integrated infrastructures to support the full experimental research cycle at experimental facilities.

Metadata is hereby seen as the key integrating factor for the different processes and data products, allowing for the easy management, discovery and access of data and tools. The Core Scientific Meta-Data Model (CSMD) developed by the Science

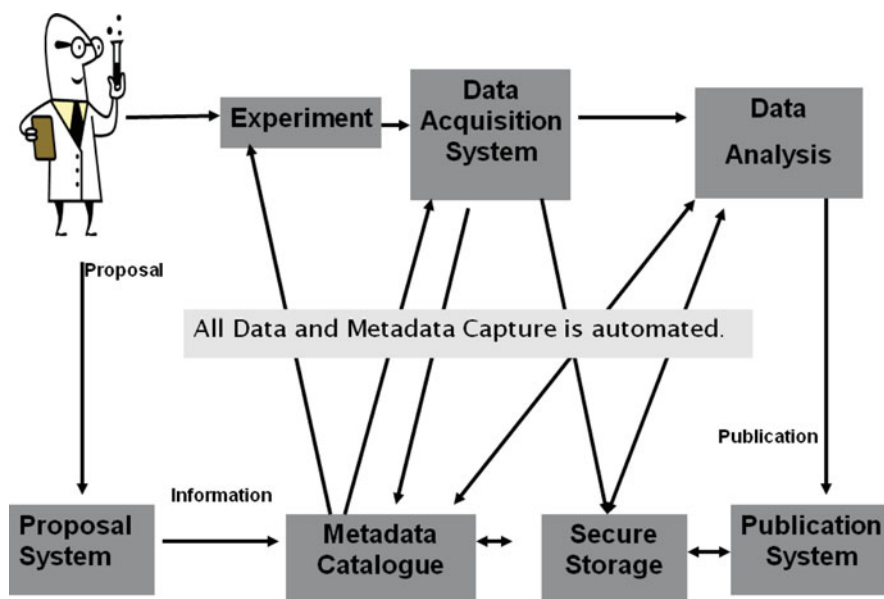


Fig. 10.2 Life cycle support at experimental facilities

and Technology Facilities Council (STFC) in the UK has hereby emerged as a *de-facto* standard [13–16]. CSMD is a study based metadata model, capturing high level information about experiments, instruments, samples and their resulting raw and derived data incl. the analysis process (see Fig. 10.3).

Its flexible structure of property lists, allows the model to be customized for any instrument type. It provides the necessary integration across investigative methods at a particular institute to support discovery and access, as well as co-analysis tasks. Scientists have furthermore the ability to link in material from related activities, which can incorporate other experiments as well as publications and presentations about the experiment. CSMD is currently used by a wide range of experimental facilities worldwide to capture and manage their scientific metadata.

Many of these institutes have developed customized infrastructure solutions for their particular facility or laboratory based around the core CSMD model. One well known example is the STFC developed integrated infrastructure for its Neutron Source ISIS, Central Laser Facility and DIAMOND Lightsource, based around the Information Catalogue (ICAT¹⁷). The software was made open source in 2008¹⁸ and was the only one available for distribution and usage by others in this field. Since its release it has been adopted by a range of other facilities in Europe, Australia, and the US. The complete infrastructure supports all process from proposal submission

¹⁷<http://www.icatproject.org/>.

¹⁸<http://code.google.com/p/icatproject/wiki/IcatMain>.

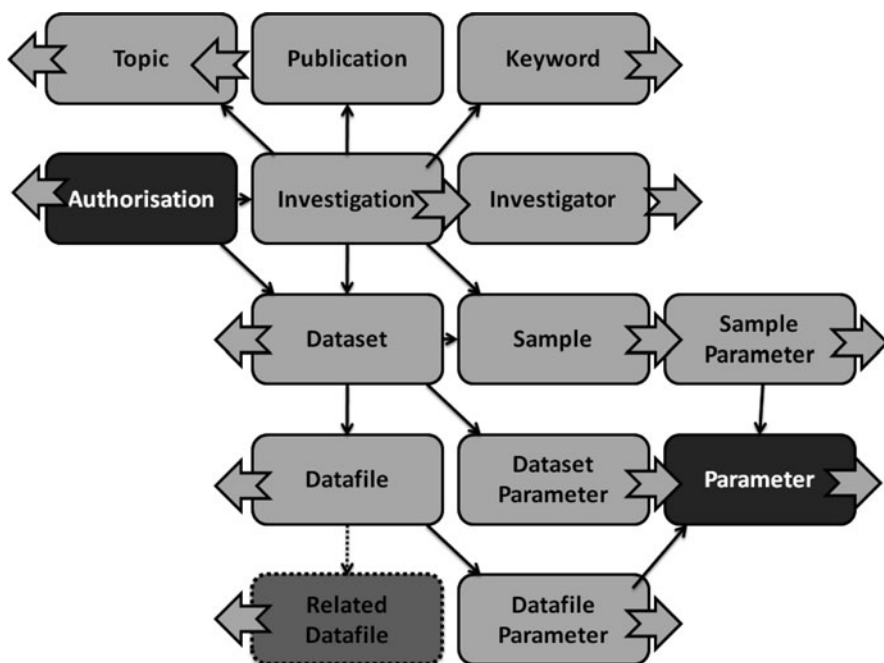


Fig. 10.3 CSMD general structure [17]

to data acquisition and distribution. A 2010 funded UK effort ‘Infrastructure for Integration in Structural Sciences’¹⁹ extended the infrastructure to support and manage the creation of derived data.

ICAT provides however only the central component of a much more complex network of services required to support the experimental process, as Fig. 10.4 below of the infrastructure set up at the UK DIAMOND facility shows.

Key challenges in such infrastructure developments are the integration of the different components, in this case facilitated through the central ICAT system and the monitoring of the correct operation and interoperation of the many different tasks. Newer infrastructure development efforts such as those at the Pacific Northwest National Laboratory (PNNL) Environmental Molecular Sciences Laboratory²⁰ have started to explore the usage of high performance workflow systems such as MeDICI²¹. Other infrastructure developments based around the CSMD model are found at the US ORNL Spallation Neutron Source (SNS), the Australian CIMA

¹⁹<http://www.ukoln.ac.uk/projects/I2S2/>.

²⁰<http://www.emsl.pnl.gov/emslweb/>.

²¹<http://dicomputing.pnnl.gov/demonstrations/medici/>.

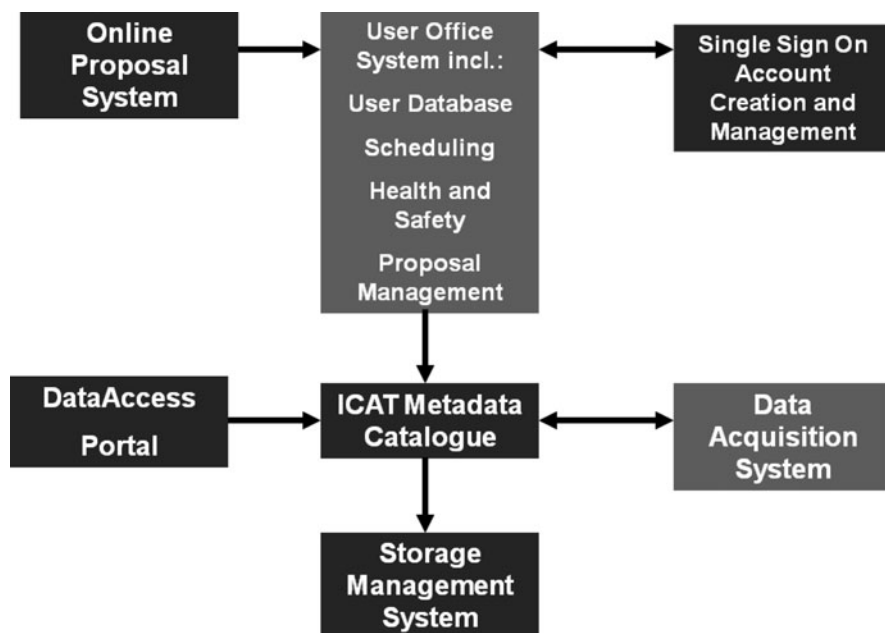


Fig. 10.4 Core data management infrastructure components

[18], Archer,²² and eCrystals (UK).²³ All of these infrastructures aim to provide an improved support for their users throughout the experimental process, delivering improved access to information and data, as well as supporting long term access and sharing of results.

3.3 *Standardization Efforts*

Within the X-ray and Neutron scattering communities, there is an emerging data format standard named NeXus²⁴ which is based upon the HDF5 self-describing data format. This is a community lead collaboration to define a suitable data standard commensurate with the needs of experimental user facilities. Undertaking such an initiative is no small task as the experimental techniques vary widely across the user facilities. Some considerations include accommodating the large variations in detector technologies and geometries, the wide variety of sample environment

²²<http://archer.edu.au/about/>.

²³<http://ecrystals.chem.soton.ac.uk/>.

²⁴http://www.nexusformat.org/Main_Page.

data to associate with the experiment data, as well as the variety of beam spectrum monitor information and beam power information. Initially the NeXus format only supported the histogram data format. NeXus was well suited to this as the data were written as a final step of creating the file. The intrinsic compress capabilities of HDF5 were employed which could result in significantly reduced file sizes. However with the advent of event based data acquisition, it was necessary to extend the NeXus format to support a list based, or streaming data format. Initially NeXus was not well suited to supporting arbitrary length data sets, though significant effort was expended to adapt NeXus to better accommodate the intrinsic unformatted nature of event data.

The data file creation occurs via a process which has been named data translation which takes raw input data files from various sources, massages, and produces a NeXus file. The granularity of the data contained within the NeXus file can be somewhat arbitrary, however for the sake of convenience; a file typically will contain the results from one data acquisition start/stop interval which is often called a “run.” The raw input data produced during a run are typically comprised of the event data list, the event pulse information (for pulsed sources such as a spallation neutron source), or the histogram data in the case of X-ray instruments where individual X-ray photons occur too rapidly to be counted individually via today’s detector technology.

The construction of the NeXus file must take into consideration the mapping of the pixel information as detector pixel locations may need to be abstracted to represent a uniform ordering rather than the order which may have resulted from producing the detector. For example, the lower left corner of the detector as viewed from the sample may be defined as the origin, however the detector as created may not define these pixels in a similar fashion. In these cases, it is necessary to re-map the pixels to a desired orientation. The mapping process places the pixels in a relative space, however it is also necessary to locate these pixels in an absolute space. To do so requires applying instrument geometry information such as path lengths, orientation angles, and measured information such as pixel spacing within the detector. Standard samples (such as silicon, diamond, or other material) can be used to fine-tune the instrument geometry information.

Experiment related metadata must also be captured and incorporated within the NeXus files. There is a wide variety of metadata to consider here and incorporate properly. The most important information pertains to the parametric conditions which the test subject, or sample, was under and called the sample environment metadata. In some cases, the sample environment data can be considerably large itself. Pressure, temperature, and magnetic field are the primary sample environment data collected. These data must be time-correlated with the measurements, particularly for event data, to best take advantage of the experiment information.

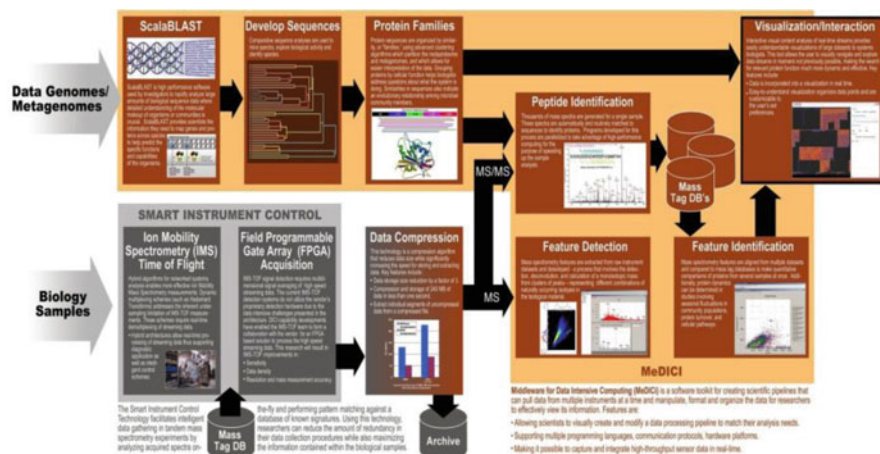


Fig. 10.5 Proteomics mass spectrometry pipeline at PNNL

3.4 Optimization of Single Experiment Analysis

The challenges in chemical imaging analysis stem from large data sets and fast data generation rates, as well as the drive to faster processing to move from post analysis to real time analysis and experimental steering. To achieve this goal the community has principally concentrated on two separate approaches, the optimization of specific analysis steps or software systems and the automation of process chains to support smother turn around.

PNNL developed for example a componentized pipeline for the control and real time analysis of proteomics mass spectrometry experiments (Fig. 10.5), using its MEDICi workflow system [19]. This is a highly standardized process at the laboratory and thus lends itself to automation via scientific workflows. The pipeline combines data intensive analytical software modules, visualization software and very large databases within a single integration framework (MeDiCi). Incoming spectra from a simulated mass spectrometer are analyzed in real time to determine the course of processing for individual samples based on comparing them to an existing results and updating the database of observed mass and time values. The same spectra are visualized within a central software component along with additional results of analytical processing. A feedback based on the results of the analytical processing is initiated back to the instrument which decides whether the samples have been fragmented already.

This capability provides a spectrum of benefits:

- Processing of already analyzed features is avoided, which allows more efficient instrument usage and reduces the amount of redundant data generation. This has positive impact in data richness and speeds the results to the end user.

- Without the smart instrument control method, experimental results of interest are usually the hardest to acquire. The described method will lead to more intelligent data gathering, which will improve analysis quality, reduce costs, and increase knowledge of the biological systems being studied.

Similarly the Medical College of Wisconsin²⁵ created e.g. an open-source cloud based environment for the analysis of proteomics results, as their own computing capacity was insufficient to serve all their users.

Complementary to these automation approaches the community has started to develop more sophisticated tools for core analysis functions. The US–UK collaboration Mantid²⁶ is working for example to consolidate the data analysis software for neutron scattering experiments by creating a flexible framework that provides a set of common services, algorithms and data objects. For robustness, the software aims to be experiment technique independent and supported on Windows, Linux, and Mac OS platforms. The goal is to provide essential functionality combined with an easy to use scripting interface, to allow scientists the ability to extend and create applications. The software is open source and the project currently has 24 active contributors with almost 600,000 lines of code written. For performance, the data objects and computation are implemented via the C++ while a python scripting interface is provided for ease of use and integration. A key feature of Mantid is its ability to read and process HDF5 based NeXus files which contain event data while heavily utilizing multi-threading to accelerate performance.

Other efforts focus on the optimization of specific analysis methods. Hereby data reduction methods can expedite the processing and subsequent feature based analysis [20, 21], including fusion, segmentation, and visualization. Similarly effective data compression at different levels of image analysis can aid the faster extraction of useful information and transfer of data to next analysis step. Segmentation algorithms must be general and intuitive for the non-expert to utilize, and to be effective in the field, the algorithms must also exhibit real-time performance. Current approaches include e.g. advanced suites of tools for segmentation based on energy minimization with Graph Cuts, Random Walks, and the Power Watershed frameworks. More work remains to be done to decrease both the computational complexity and the memory footprint of these approaches. Feature detection is another critical component in the analysis process, allowing to emphasize feature of interest by removing disturbing artifacts and background noise [22, 23]. In general, much effort still needs to be expended to address the efficiency of the analysis algorithms themselves, many of which remain to date sequential algorithms, which are not adapted to meet the needs of the data intensive requirements of experimental analysis. Parallelized algorithms are crucial to real time measurement, analysis, and control of scientific experiments. Initial efforts at e.g., Lawrence

²⁵<http://www.genomeweb.com/informatics/mcw-insilicos-enable-open-source-proteomics-tools-data-analysis-cloud>.

²⁶Mantid Project home page: http://www.mantidproject.org/Main_Page.

Berkley National Laboratory and ORNL are focusing on the usage of parallel algorithms and high performance computing to speed up the analysis algorithms themselves [24].

3.5 Data Exchange, Sharing, and Publication

At the same time that technology transformed how we do research and analysis, science also transformed with whom we work, research today is much more collaborative, international and interdisciplinary than it was 50 years ago. Geographically dispersed collaborations are common practice today, even across several continents and the single researcher or closed local collaborations are a rarity nowadays [25, 26]. It is clear that with the advent of a much deeper understanding of scientific subjects and increasingly complex experimental and computational technologies a strong individual specialization, not only along the lines of scientific topics but also research methodologies has taken place [27, 28]. On the other hand societal problems drive funders to encourage science to help with the solution for much more complex challenges requiring interdisciplinary (integration of different science domains) and multidisciplinary (several disciplines make a separate contribution) projects or borrowing (usage of technologies from a different discipline), thus a much broader, non domain specific scientific knowledge and information exchange [29]. This exchange forms the basis for the more important collaborative tasks of co-analysis and visualization of results across techniques and disciplines.

The general working practices around the sharing of research results have however not changed much over the past centuries, research publications are still the main sources of information exchange. Unfortunately publications have certain limitations in conveying comprehensive information on a particular subject, there is the restriction in length and thus detail, as well as that its main purpose is to convey ones point of view rather than necessarily a comprehensive, objective representation of all facts [30–33]. Publications thus provide at best a very coarse and high level summary of the research work undertaken by the authors, but are not suitable in supporting co-analysis tasks. The associated raw and derived data would be rich source of supporting information, in particular if coupled with the appropriate metadata and documented scientific workflows [34].

In recognition of the desire by the research community to get access not only to the summary of a research project, but also the underpinning data, more publishers today require from their authors that they share their raw and derived data by depositing it into publicly accessible archives or by providing it on request. However, recent studies have shown [35, 36] that few authors comply with the latter requirement, and only the enforced deposition before publication seems to work. This seems to indicate a continued reluctance to share in-depth research results with the general research community. Nevertheless a growing awareness of the value of the produced data as research record in its own right has given

rise to the creation of a large number of new institutional and community data collections [37] and an exponential growth of the existing ones [38]. The drivers for the creation of these collections are usually organizations and funders, rather than researchers themselves, this is demonstrated by the low data deposition rates even in highly regulated data publication subjects such as Crystallography were only around 20% of all determined structures are publicly accessible, a mere 1.3% of all known Chemical compounds [39].

The advent of citable data publications is however slowly turning the tide in a number of research communities, in particular organizations such as DataCite²⁷ work to increase the acceptance of research data as legitimate, citable contributions to the scientific record. The ORCID²⁸ collaboration on the other hand, works on removing the ambiguity in attributing such research records reliably to specific scientists.

While researchers may still be reluctant in many fields to share their data more globally, it is a core necessity for them to share their data and progress with their fellow collaborators. In 2008 a NSF workshop on “*New Models for scholarly Communication in Chemistry*” investigated the merits of introducing new web based methods of collaboration and communication into chemistry and thus experimental sciences. Whilst methods such as semantic web, semantic publishing, open notebook science and data publishing were seen as embryonic at the time and had not yet found a broader user base, their undoubted potential to enhance scientific communication was clearly identified [40]. Since then technology has progressed and a number of interesting developments have emerged in particular from the former e-Science community. The international LabTrove²⁹ development combines integrated data management infrastructures for experimental sciences with online blogging to create a smart research framework. LabTrove integrates a number of key developments: Electronic Laboratory Notebook, Collaborative research support through MyExperiment,³⁰ an experimental ontology and a blog factory [41]. Similarly the PNNL development Velo [42] combines a classical content management system (Alfresco) with a semantic media Wiki and the collaborative analytical toolbox (CAT)³¹ to provide project based collaborative environment for discussions, data sharing and co-analysis. Velo is currently used by a wide range of different communities including a number of experimental groups.

²⁷<http://www.datacite.org/>.

²⁸<http://www.orcid.org/>.

²⁹<http://www.labtrove.org/>.

³⁰<http://www.myexperiment.org/>.

³¹<http://omics.pnl.gov/software/CAT.php>.

4 Future Directions

A key medium term challenge is the routine co-analysis of scientific results and the improvement of analysis tools in general to move towards more sophisticated community tools that are suitable for both high data volumes and real or near-real time analysis. Initial efforts are emerging to build the necessary infrastructure and tools that would offer such capabilities. In the longer term data intensive analysis for experimental facilities should become an integral part of a more general data intensive environment that combines both experimental and computational approaches.

4.1 *Co-Analysis Across Different Investigative Methods*

Today's scientific challenges are complex and usually require the integration of a wide range of complementary expertise to be addressed successfully. Research results from a wide range of experimental imaging technologies, ranging from nano to macroscale, need to be brought together to form a coherent synergistic picture. At present, however, scientists are usually only familiar with a very limited range of experimental technologies. Each of these different technologies currently requires in-depth domain knowledge to enable the user to use the technique correctly and to be able to interpret the results correctly. Each scientist can therefore only make use of a very limited palette of experimental technologies. They are thus limited in their ability to synthesize and connect their own research with the work of others, who are investigating the same or related topics, but with different experimental technologies. The ability to go beyond such limitations through a clear understanding of what each of these technologies delivers in terms of scientific insights, and the ability to synthesize results across a wide spectrum of imaging technologies would be a powerful catalyst for the quality and pace at which scientific research and discovery can be carried out. In addition, it would be crucial for the faster exploitation of those results by industry and academics.

Image informatics is a developing interdisciplinary field of research that encompasses computer science, statistics, engineering, mathematics, information science, as well as the natural sciences. The primary challenge is to maximize experimental outcomes by enabling the correct end to end analysis. If an important bit of data or metadata is lost or converted into the wrong form for preservation, it is gone and expensive experiments do not reach their potential or have to be repeated. The focus of current research in PNNL's Chemical Imaging Initiative³² is to define a framework for chemical imaging co-analysis (Fig. 10.6). This framework

³²<http://www.pnl.gov/science/research/chemicalimaging/>.

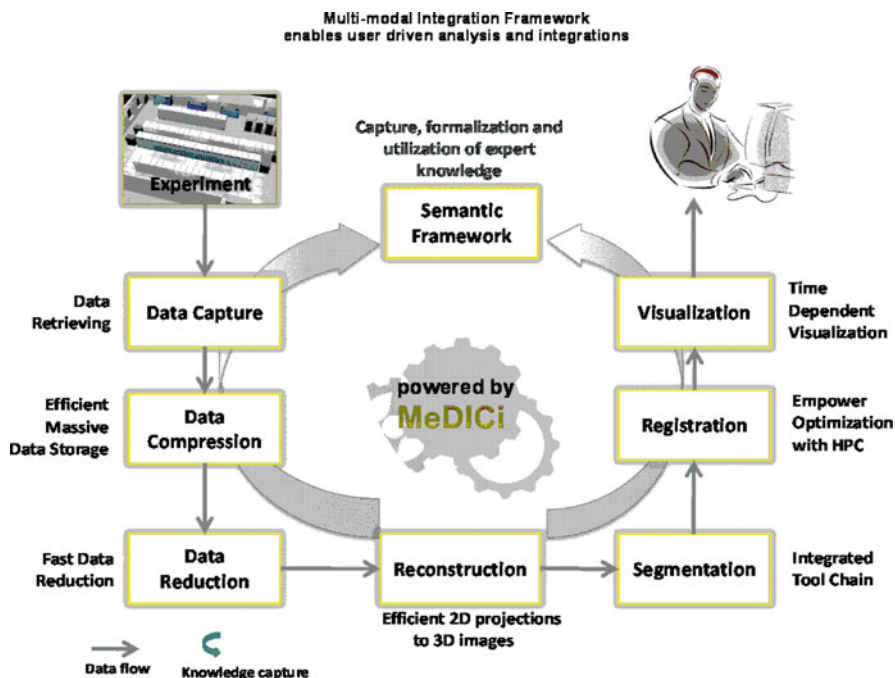


Fig. 10.6 High level framework overview

will necessarily include capabilities for data preservation, data description, data management, and data analysis. There is currently no suitable framework available or under development worldwide that the authors are aware of that appropriately handles the multitude of chemical imaging methodologies and the petabytes of data expected to be generated. The routine co-analysis of experimental results from different imaging technologies has so far not been addressed. The proposed framework will bring together a range of existing research concepts in the areas of semantic mapping and rules, workflows, and core technologies for the capture, analysis and integration of experimental data, integrate and develop these further to create this unique capability.

The workflow architecture and the semantic framework will ensure the coherence of the knowledge capture, exploitation and usage by the different components. The framework raises the integration needs with emerging requirements on functions, data types, semantics and real-time properties of the workflow to be addressed at an overarching level. The group anticipates that the data exchange between imaging technologies will be complex and intensive (petabytes of data to be generated), with the rapid growths of data sets spanning different spatial and temporal scales. In response to the challenges of data intensive integration of imaging technologies, this architecture is being built by leveraging PNNL's MeDICi (Middleware for Data Intensive Computing). MeDICi is a middleware platform for building complex, high

performance analytical applications. It has been proven efficient and successful as the communication backbone and execution environment for scientific and analytic domain applications. Leveraging MeDICi, the group aims to explore the transformations needed to take data from one technique, tool, or application and feed it into another during the workflow execution. The project focuses on identifying the intrinsic linkage of the imaging technologies and understanding data characteristics.

The semantic component of the framework will consist of four components: Characterization, Relation, Analysis, and Representation. Basic concepts for these areas have been developed and tested for a range of projects [43–45]. However, these were never applied collectively, nor integrated to capture domain knowledge in an easily usable form.

Starting with six key imaging technologies initially chosen for the initiative, the group will develop formal characterizations of the methods, instruments, samples, analysis processes, and data products associated with each of these, detailing in particular what each method contributes to the overall domain knowledge. Furthermore, they will determine how each of these methodologies relates to the others (for example A refines B, A complements B by adding X), thus building a formalized topology of the methods, their contribution, and constraints. Based on these initial characterizations they formalize their functionalities, so that it can be extended to other techniques and utilized by a wider community.

A further enabling technology identified for the success of the framework is the ability for distributed analysis. The instrumentation used to collect experimental data is expected to continue to improve in resolution and size, thus resultant data sets can grow into the multi petabyte range. Furthermore the facilities housing these results will be geographically distributed. While it is possible today to transfer a few terabytes of data across thousands of miles in a day, poor and unpredictable data transfer rates are the norm over long distances on wide area networks. If the performance of long-distance file transfers cannot be assured, the best alternative is to minimize the quantity of data that must be transferred. Failing that, the computation must be brought to the data. The initiative will therefore investigate new analysis methods that can work across distributed data sets.

In this light there has also been a recent proposal to establish a user facility network (UFnet) which would facilitate inter-facility data movement and management [46]. An initial focus would be the routine integration of multi-technique data from X-ray and neutron sources. Tech-X Corporation's open source tool Orbiter³³ provides in this context for example:

- Secure User and Management Interface: Users, managers, and resource providers demand a rich environment of tools, tutorials, documentation, and customizable interfaces that can be accessed from Internet capable mobile phones, laptops, and workstations.

³³<https://orbiter.txcorp.com>.

- Scalable Virtual Organization and Community Management: We envision not only capitalizing on role-based infrastructures but also providing federated community identity management capabilities. It is essential that a scalable management infrastructure provide the ability for DOE stakeholders to audit and organize their personnel usage by project, department, or service type.
- Dynamic User-Centric Compute and Storage Resource Status: Up-to-date resource status, state, and load is required to dynamically scale enterprise service infrastructures to meet the stakeholder throughput, storage, and bandwidth requirements for all resources.
- Reliable Resource Configuration and Management: Efficient and reliable infrastructure application deployment and configuration management provides the feedback necessary for optimizing the deployed applications and services required to differentiate the Orbiter UFnet production and productivity systems.
- Easy to use access to HPC mechanisms, via thick and thin clients supporting Service Oriented Architecture (SOA) based services consist of standards-based components that are reusable and extensible for accessing high performance computing, data and computational grid infrastructure, and cluster-based resources easily from a user configurable interface.
- A prototype network node services to enable off-line and online simultaneous multi-technique experiment and analysis for X-ray scattering at APS and neutron scattering at SNS is shown in Fig. 10.7.

4.2 Long Term Perspective

For the future it is hoped that data will work for scientists rather than scientists working for their data – network, data, computing infrastructure, and software will be synergistically integrated to better enable collaborative pursuit of scientific discoveries resulting from experiments performed at user facilities. Data management and analysis would hereby be a central component of any such solution and data issues would be considered as an integral component of any system design [47]. A range of forward looking white papers on data intensive science have discussed the issues involved in establishing such wide reaching infrastructures and proposed options for the way forward [48–52]. Each of these is focused on seamless access to research data and the provision of advanced analysis capabilities.

Open Access and Data Curation (long term preservation for reuse) issues have long driven the development of standards, methods and infrastructures for data intensive science in Europe. The 2008 update to the roadmap of the European Strategy Forum on Research Infrastructures (ESFRI) lists for the first time the need not only for leading edge experimental and computational facilities to drive future scientific progress, but also adds the importance of an underpinning e-infrastructure consisting of integrated communication networks, distributed grids, high performance computing, and digital repositories components. ESFRI states further that data in their various forms (from raw data to scientific publications)

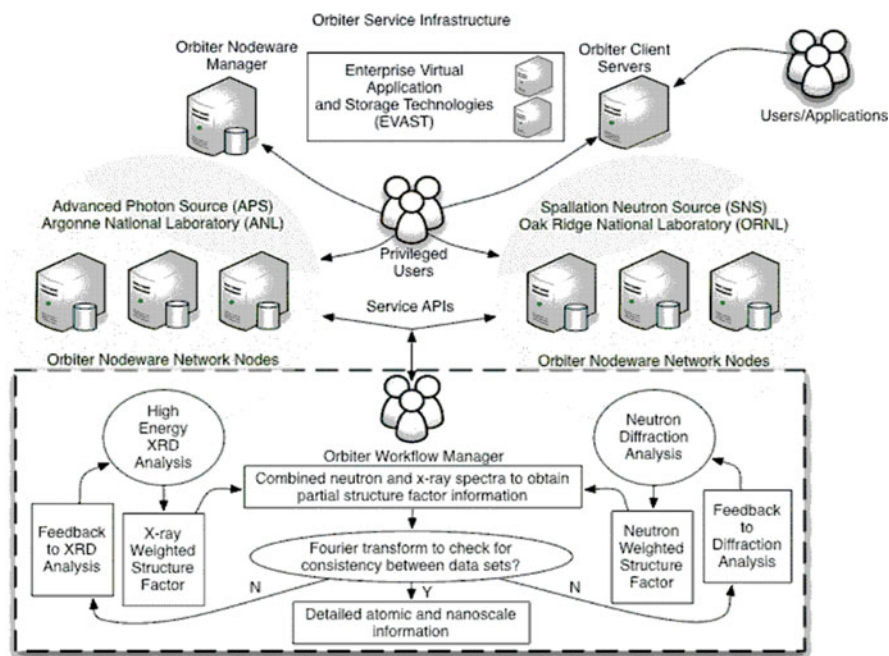


Fig. 10.7 User facility network prototype science case study features and functionality

will need to be stored, maintained and made available and openly accessible to all scientific communities. They are placing a new emphasize on digital repositories as places to capture and curate the scientific data both for the good of science and the economy. Intellectual and Technological progress in these areas has particularly been driven by centers of excellence, large scale long term infrastructure projects and organizations with visionary leadership and an in-depth understanding of data intensive sciences. Key examples for international centre's and projects are: UK Data Curation Centre,³⁴ US SciDAC SDM centre, the Earth Systems Grid³⁵ and its international Partners, e-Infrastructure for Large Scale Experimental Facilities [13] and the Biomedical Informatics Research Network (BIRN³⁶). These projects have clearly demonstrated the potential of data intensive science technologies; however as the report 'Data-Intensive Research Theme' [49] notes '*Current strategies for supporting it demonstrate the power and potential of the new methods. However, they are not a sustainable strategy as they demand far too much expertise and help in addressing each new data-intensive task*'. This and other recent publication [48,51] clearly show the community consensus that more generalized, easy to use solutions

³⁴<http://www.dcc.ac.uk/>.

³⁵<http://www.earthsystemgrid.org/>.

³⁶<http://www.birncommunity.org/>.

need to be developed to make a more wide spread use of these basic data intensive technologies possible. Thought leaders are also pointing out, that while the current developments of infrastructure surrounding the management of data continue to be important, it is time to go beyond these basic approaches and focus on the data itself – developing the means to transform data into an infrastructure in its own right. In response the European Union announced in 2010 a high level funding opportunity to develop new seamless infrastructure demonstrators across a wide range of computational and experimental resources, with the first projects set to start in late 2011.

In the US the National Science Foundation has proved to be a driving force for change, by requiring structured data management plans from all grant applicants. Furthermore the NSF is in regular discussions with its European counter parts to explore the potential for a harmonization of policies and infrastructures. A recent NSF-OCI Task Force on Data and Visualization recommended to [53]:

- Identify and share best-practices for the critical areas of data management
- Effectively and securely offer data services/access to various stakeholder communities
- Associate scientific publications with the underlying data and software assets (to improve the reproducibility of science)

5 Conclusions

Experimental research methods can offer fundamental insights, gained through a large variety of investigative methods, to help address pressing, complex scientific challenges. Hereby direct imaging methods are used to probe structure, properties, and function of objects from single elements to whole communities, helping to develop an atomistic understanding of scientific issues. Advances in the underlying experimental technologies have lead to an exponential growth in the volumes, variety and complexity of data derived from such methodologies, making experimental science a very data intensive field of science. This exceptional growth in data volumes and complexity has presented researchers with significant challenges, foremost how to effectively analyze the results of their research both for single experiments and increasingly across different investigative methods. Key issues are:

- Algorithms are often unable to handle the increasing volumes and diversity of the data either at all or in a timely fashion
- The community requirement for real time analysis cannot be met with present solutions
- While it has been acknowledged that scientific discovery, like medical diagnosis of a patient's condition, require integration of inputs and findings from a number of sources there are no routine co-analysis techniques available to the community

Furthermore experimental analysis relies heavily on the availability of a supporting eco-system of data management, movement, security, and collaboration services to be successful.

Community efforts so far have largely concentrated on the improvement of data management eco-system at experimental facilities by developing policies, standards, and integrated infrastructures. The optimization of single experiment analysis through improved methods and automated analysis pipelines has been a more recent focus of the community's research efforts, with a number of exemplary successes in the area of automation. Only a few small developments are currently emerging in the field of collaboration support for experimental sciences. Initial research work is emerging focused on building the necessary infrastructure and tools to support routine co-analysis of scientific results; however, these projects are still in their infancy and so this domain is seen as a fertile growth area with many research challenges still ahead.

Overall, while progress is being made on the development of supportive data management eco-systems, the key data intensive analysis challenges for experimental facilities remain. There is a critical lack of analytical methods that can routinely and reliably handle the growing volume and diversity of data, support real time and co-analysis. Image informatics the interdisciplinary field of research that encompasses computer science, statistics, engineering, mathematics, information science and natural sciences, as well as data intensive science research itself would seem to offer the most promising approaches to solving these analysis challenges and enable the crucial progress for experimental sciences.

Acknowledgements S.D.M. acknowledges that the research at Oak Ridge National Laboratory's Spallation Neutron Source was sponsored by the Scientific User Facilities Division, Office of Basic Energy Sciences, U. S. Department of Energy.

S.D.M and J.W.C. acknowledge that the submitted manuscript has been co-authored by a contractor of the U.S. Government under Contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

J.W.C. acknowledges that this material is based upon work supported by the National Science Foundation under Grant No. 050474. This research was supported in part by the National Science Foundation through TeraGrid resources provided by the Neutron Science TeraGrid Gateway.

References

1. National Research Council. *Visualizing Chemistry: The Progress and Promise of Advanced Chemical Imaging*, The National Academies Press, Washington, DC, 2006.
2. Basic Energy Science Advisory Committee, Subcommittee on Facing Our Energy Challenges in a New Era of Science, "Next Generation Photon Sources for Grand Challenges in Science and Energy", Technical Report, U.S. Department of Energy, May 2009.
3. F. Maia, P. van der Meulen, A. Ourmazd, I. Vartanyes, G. Bortel, K. Wrona, M. Altarelli, G. Huldt, D. Larsson, R. Abela, V. Elser, T. Ekeberg, K. Cameron, D. van der Spoel, H. Kono, F. Wang, P. Thibault, and A. Mancuso, "Data Analysis and its needs @ European Xfel". Presentation SPB-Workshop 2008 Working Group 3. http://www.xfel.eu/events/workshops/2008/spb_workshop_2008/ (accessed May 6th 2011)

4. C. Southan and G. Cameron, "Beyond the Tsunami: Developing the Infrastructure to Deal with Life Sciences data" In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, 2009, Microsoft Research.
5. C. Goble and D. De Roure, "The Impact of Workflow Tools on Data-centric Research" In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, 2009, Microsoft Research.
6. K. Alapaty, B. Allen, G. Bell, D. Benton, T. Brettin, S. Canon, R. Carlson, S. Cotter, S. Crivelli, E. Dart, V. Dattoria, N. Desai, R. Egan, J. Flick, K. Goodwin, S. Gregurick, S. Hicks, B. Johnston, B. de Jong, K. Kleese van Dam, M. Livny, V. Markowitz, J. McGraw, R. McCord, C. Oehmen, K. Regimbal, G. Shipman, G. Strand, B. Tierney, S. Turnbull, D. Williams, and J. Zurawski, "BER Science Network Requirements", Report of the Biological and Environmental Research Network Requirements Workshop, April 29 and 30, 2010, Editors E. Dart and B. Tierney, LBNL report LBNL-4089E, October 2010.
7. B.F. Jones, S. Wuchty, and B. Uzzi, "Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science" in *Science Express* on 9 October 2008, *Science* 21 November 2008: Vol. 322. no. 5905, pp. 1259–1262
8. E. Yang, "Martin Dove's RMC Workflow Diagram", a supplementary requirement report, Work Package 1, November 2009 – June 2010, JISC I2S2 project, July 2010, available at: <http://www.ukoln.ac.uk/projects/I2S2/documents/ISIS%20RMC%20workflow.pdf>
9. E. Dart and B. Tierney, "BES Science Network Requirements – Report of the Basic Energy Sciences Network Requirements Workshop Conducted September 22 and 23, 2010".
10. S.D. Miller, A. Geist, K.W. Herwig, P.F. Peterson, M.A. Reuter, S. Ren, J.C. Bilheux, S.I. Campbell, J.A. Kohl, S.S. Vazhkudai, J.W. Cobb, V.E. Lynch, M. Chen, J.R. Trater, B.C. Smith, T. Swain, J. Huang, R. Mikkelsen, D. Mikkelsen, and M.L. Green, "The SNS/HFIR Web Portal System – How Can it Help Me?" 2010 J. Phys.: Conf. Ser. 251 012096. doi:10.1088/1742-6596/251/1/012096.
11. Federal Information Processing Standards Publication – FIPS PUB 199, "Standards for Security Categorization of Federal Information and Information Systems" February 2004.
12. Scientific Data Management (SDM) for Government Agencies: Report from the Workshop to Improve SDM. "Harnessing the Power of Digital Data: Taking the Next Step. June 29-July 1, 2010.
13. D. Flannery, B. Matthews, T. Griffin, J. Bicarregui, M. Gleave, L. Lerusse, S. Sufi, G. Drinkwater, and K. Kleese van Dam, "ICAT: Integrating data infrastructure for facilities based science". Proc. 5th IEEE International Conference on e-Science (e-science 2009), Oxford, UK, 09–11 Dec 2009
14. S. Sufi, B. Matthews, and K. Kleese van Dam. (2003) *An Interdisciplinary Model for the Representation of Scientific Studies and Associated Data Holdings*. UK e-Science All Hands meeting, Nottingham, 02–04 Sep 2003
15. S. Sufi and B.M. Matthews. (2005) *The CCLRC Scientific Metadata Model: a metadata model for the exploitation of scientific studies and associated data*. In Contributions in Knowledge and Data Management in Grids, eds. Domenico Talia, Angelos Bilas, Marios Dikaiaikos, CoreGRID 3, Springer-Verlag, 2005.
16. E. Yang, B. Matthews, and M. Wilson, "Enhancing the Core Scientific Metadata Model to Incorporate Derived Data," eScience, IEEE International Conference on, pp. 145–152, 2010 IEEE Sixth International Conference on e-Science, 2010
17. B. Matthews, "Using a Core Scientific Metadata Model in Large-Scale Facilities". Presentation at 5th International Digital Curation Conference (IDCC 2009), London, UK, 02–04 Dec 2009
18. I.M. Atkinson, D. du Boulay, C. Chee, K. Chiu, T. King, D.F. McMullen, R. Quilici, N.G.D. Sim, P. Turner, and M. Wyatt, "CIMA Based Remote Instrument and Data Access: An Extension into the Australian e-Science Environment." *Proceedings of IEEE International Conference on e-Science and Grid Computing (e-Science 2006)* Amsterdam, The Netherlands, December 2006.
19. I. Gorton, A. Wynne, Y. Liu, and J. Yin, "Components in the Pipeline," *IEEE Software*, vol. 28, no. 3, pp. 34–40, May/June 2011, doi:10.1109/MS.2011.23

20. D. Li, M. Tschoopp, X. Sun and M. Khaleel, Comparison of reconstructed spatial microstructure images using different statistical descriptors. Submitted to *Computational Materials Science*
21. D. Li Application of chemical image reconstruction on materials science and technology, accepted by Proceeding of 2011 World Congress of Engineering and Technology, IEEE, and will present the paper in October 2011
22. L.M. Kindle, I.A. Kakadiaris, T. Ju, and J.P. Carson (2011) A semiautomated approach for artefact removal in serial tissue cryosections. *Journal of Microscopy*. 241(2):200–6.
23. J.P. Carson, D.R. Einstein, K.R. Minard, M.V. Fanucchi, C.D. Wallis, and R.A Corley (2010) High resolution lung airway cast segmentation with proper topology suitable for computational fluid dynamic simulations. *Computerized Medical Imaging and Graphics*. In Press.
24. M. Hohn, G. Tang, G. Goodyear, P.R. Baldwin, Z. Huang, P.A. Penczek, C. Yang, R.M. Glaeser, P.D. Adams, and S.J. Ludtke, “SPARX, a new environment for Cryo-EM image processing” in *J Struct Biol*. **157**, 47–55, 2007
25. B.F. Jones, S. Wuchty, and B. Uzzi, 2008. ‘Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science’ in *Science Express* on 9 October 2008, *Science* 21 November 2008: Vol. 322. no. 5905, pp. 1259–1262
26. R. Guimera, B. Uzzi, J. Spiro, and L.A.N. Amaral, 2005. ‘Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance’ in *Science*, 308, 697 (2005).
27. M. Pianta and D. Archibugi, 1991. ‘Specialization and size of scientific activities: A bibliometric analysis of advanced countries’ in *Scientometrics* Volume 22, Number 3/November, 1991
28. W. West and P. Nightingale, 2009. ‘Organizing for innovation: towards successful translational research’ in *Trends in Biotechnology*, Volume 27, Issue 10, 558–561, 17 August 2009
29. Committee on Facilitating Interdisciplinary Research, National Academy of Sciences, National Academy of Engineering, Institute of Medicine. 2004. ‘The Drivers for Interdisciplinary Research’ in *Facilitating interdisciplinary Research* p 26–40, 2004
30. D. Shotton, K. Portwin, G. Klyne, and A. Miles, 2009. ‘Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article’ in *Publication Library of Science Computational Biology*. 2009 April; 5(4).
31. A. de Waard, L. Breure, J.G. Kircz, and H. van Oostendorp, 2006. ‘Modeling rhetoric in scientific publication’ in *Proceedings of the International Conference on Multidisciplinary Information Sciences and Technologies*, pp 1–5, InSciT2006; 25–28 October 2006; Merida, Spain. <http://www.instac.es/inscit2006/papers/pdf/133.pdf>.
32. T. Kuhn, 1962. *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1962)
33. B. Latour, 1987. ‘Science in Action’ in *How to Follow Scientists and Engineers through Society*, Cambridge, Ma.: Harvard University Press, 1987.
34. C. Goble and D. deRoure, 2009. “The impact of Workflow tools on data-centric research” In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, 2009, Microsoft Research.
35. C.J. Savage and A.J. Vickers (2009) Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. *PLoS ONE* 4(9): e7078. doi:10.1371/journal.pone.0007078.
36. J.M. Wicherts, D. Borsboom, J. Kats, and D. Molenaar, 2006. ‘The poor availability of psychological research data for reanalysis’ in *American Psychologist* 61: 726–728.
37. D. De Roure, C. Goble, S. Aleksejevs, S. Bechhofer, J. Bhagat, D. Cruickshank, D. Michaelides, and D. Newman, 2009. ‘The myExperiment Open Repository for Scientific Workflows’ in: *Open Repositories 2009*, May 2009, Atlanta, Georgia, US. (Submitted).
38. C. Southan and G. Cameron, 2009. “Beyond the Tsunami: Developing the Infrastructure to Deal with Life Sciences data” In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, 2009, Microsoft Research.
39. S. Coles and L. Carr, 2008. ‘Experiences with Repositories & Blogs in Laboratories’ in *Proceedings of: Third International Conference on Open Repositories 2008*, 1–4 April 2008, Southampton, United Kingdom.
40. T. Velden and C. Lagoze, The Value of new Communication Models for Chemistry, White Paper 2009, eCommens@Cornell, <http://hdl.handle.net/1813/14150>.

41. J.D. Blower, A. Santokhee, A.J. Milsted, and J.G. Frey, BlogMyData: a Virtual Research Environment for collaborative visualization of environmental data. All Hands Meeting 2010, Cardiff UK 13–16 Sep 2010 <http://eprints.soton.ac.uk/164533/>.
42. I. Gorton, C. Sivaramakrishnan, G. Black, S. White, S. Purohit, M. Madison, and K. Schuchardt, 2011. Velo: riding the knowledge management wave for simulation and modeling. In *Proceeding of the 4th international workshop on Software engineering for computational science and engineering* (SECSE '11). ACM, New York, NY, USA, 32–40.
43. L.E.C. Roberts, L.J. Blanshard, K. Kleese Van Dam, L. Price, S.L. Price, and I. Brown, Providing an Effective Data Infrastructure for the Simulation of Complex Materials. Proc. *UK e-Science Programme All Hands Meeting 2006* (AHM 2006).
44. A.M. Walker, R.P. Bruin, M.T. Dove, T.O.H. White, K. Kleese van Dam, and R.P. Tyer. Integrating computing, data and collaboration grids: the RMCS tool. *Philosophical Transactions of The Royal Society A* 367 (1890) 1047–1050 (2009) [doi:10.1098/rsta.2008.0159]
45. A. Woolf, B. Lawrence, R. Lowry, K. Kleese van Dam, R. Cramer, and M. Gutierrez. Data integration with the Climate Science Modelling Language Proc. *European Geosciences Union General Assembly 2005, Vienna, Austria, 24–29 Apr 2005*, Geophysical Research Abstracts, Volume 7, 08775, 2005 (2005), *Fourth GO-ESSP meeting, RAL, UK, 06–08 Jun 2005, Workshop on Grid Middleware and Geospatial Standards for Earth System Science Data, NESC workshop, Edinburgh, Scotland, 06–08 Sep 2005*.
46. S.D. Miller, K.W. Herwig, S. Ren, S.S. Vazhkusai, P.R. Jemian, S. Luitz, A.A. Salnikov, I. Gaponenko, T. Proffen, P. Lewis, and M.L. Green, “Data Management and Its Role in Delivering Science at DOE BES User Facilities – Past, Present, and Future.
47. J. Ahrens, B. Hendrickson, S. Miller, R. Ross, and D. Williams, “Data Intensive Science in the Department of Energy” October 2010, LA-UR-10-07088.
48. K. Koski, C. Gheller, S. Heinzel, A. Kennedy, A. Streit, and P. Wittenburg. Strategy for a European Data Infrastructure: White Paper. Technical report, Partnership for Advanced Data in Europe (PARADE), September 2009.
49. M. Atkinson, M. Kersten, A. Szalay, and J. van Hemert. Data Intensive Research Theme. NESC Technical Report, May 2010.
50. J. Wood, T. Anderson, A. Bachem, C. Best, F. Genova, D. Lopez, W. Los, M. Marinucci, L. Romary, H. Van de Sompel, J. Vigen, P. Wittenburg, D. Giarretta, R.L. Hudson. Riding the Wave – How Europe can gain from the rising tide of scientific data, October 2010.
51. J. Ahrens, B. Hendrickson, G. Long, S. Miller, R. Ross, and D. Williams. Data Intensive Science in the Department of Energy, October 2010.
52. K. Kleese van Dam, T. Critchlow, J. Johnson, I. Gorton, D. Daly, R. Russell, and J. Feo. The Future of Data Intensive Science Experimenting in Data - Across the Scales, Across Technologies, Across the Disciplines. PNNL White Paper, November 2010. <https://sites.google.com/site/dataintensivesciencecommunity/home>
53. D. Atkins, T. Detterich, T. Hey, S. Baker, S. Feldman, and L. Lyon, NSF-OCI Task Force on Data and Visualization, March 7, 2011.
54. P. Rich, “Infrastructure III”, I/O Tutorial, An Advanced Simulation & Computing (ASC) Academic Strategic Alliances Program (ASAP) Center at The University of Chicago, 2009, http://flash.uchicago.edu/website/codesupport/tutorial_talks/June2009/IO_tutorial.pdf (accessed May 6th 2011)
55. Scientific Grand Challenges – Discovery in Basic Energy Sciences: the Role of Computing at the Extreme Scale, Report of DOE workshop, August 13–15, Washington DC.
56. B. Fultz, K.W. Herwig, and G.G. Long, “Computational Scattering Science 2010”, Workshop held at Argonne National Laboratory July 7–9 2010. Workshop report. <http://neutrons scattering.org/2011/01/computational-scattering-science>