PANDATA EUROPE
Open Data Infrastructure

PANDATA-ODI

Capacities - Research Infrastructures
Combination of Collaborative Project and Coordination and Support Action:
Integrated Infrastructure Initiative (I3)
INFRA-2011-1.2.2: Data infrastructures for e-Science

Name of the coordinating person: **Dr Juan Bicarregui**

List of participants:

Participant number	Participant organisation name	Participant short name	Country
1 (Coordinator)	Science Technology Facility Council	STFC	UK
2	European Synchrotron Radiation Facility	ESRF	International Organisation, F
3	Institut Laue Langevin	ILL	International Organisation, F
4	Diamond Light Source Ltd	DIAMOND	UK
5	Paul Scherrer Institut	PSI	CH
6	Deutsches Electronen Synchrotron	DESY	D
7	Sincrotrone Trieste S.C.p.A.	ELETTRA	I
8	Soleil Synchrotron	SOLEIL	F
9	Cells - Alba	ALBA	SP
10	Helmholtz Zentrum Berlin	HZB	D
11	Commissariat à l'energie atomique, Laboratoire Léon Brillouin	CEA	F

Table of Contents

Proposal Abstract

The PaN-data collaboration brings together eleven large multidisciplinary Research Infrastructures which operates hundreds of instruments used by over 30,000 scientists each year. They support fields as varied as physics, chemistry, biology, material sciences, energy technology, environmental science, medical technology and cultural heritage. Applications are numerous, for example, crystallography can reveal the structures of viruses and proteins important for the development of new drugs; neutron scattering can identify stresses within engineering components such as turbine blades, and tomography can image microscopic details of the 3D-structure of the brain. Commercial users include the pharmaceutical, petrochemical and microelectronic industries.

PaNdata-ODI will develop, deploy and operate an Open Data Infrastructure across the participating facilities with user and data services which support the tracing of provenance of data, preservation, and scalability through parallel access. It will be instantiated through three virtual laboratories supporting powder diffraction, small angle scattering and tomography.

1 SCIENTIFIC AND/OR TECHNICAL QUALITY, RELEVANT TO THE TOPICS ADDRESSED BY THE CALL

1.1 Concept and objectives

1.1.1 Background

PaNdata Open Data Infrastructure is a proposal to construct and operate a sustainable data infrastructure for European Photon and Neutron laboratories. This will enhance all research done in the neutron and photon communities by making scientific data accessible allowing experiments to be carried out jointly in several laboratories.

Formed in 2008, the PaNdata collaboration currently brings together eleven major world class European Research Infrastructures to construct and operate a common data infrastructure for the European Neutron and Photon large facilities (See www.pandata.eu). In 2010, the consortium began a Support Action which is focusing on standardisation activities in the areas of: data policy, user information exchange, scientific data formats, interoperation of data analysis software, and integration and cross-linking of research outputs. These standards form the baseline for PaNdata ODI and will ensure that the research and development activities deliver outputs that can readily be deployed into common services which integrate data across the consortium to create a fully integrated, pan-European, research data infrastructure supporting numerous scientific communities across Europe.

Scientifically, neutron and photon laboratories are complementary research facilities, often focussing on different aspects of the wide research spectrum covered by these facilities. They support experiments in many scientific fields as varied as physics, chemistry, biology, material sciences, energy technology, environmental science, medical technology and even cultural heritage investigations. Industrial applications are growing, notably in the fields of pharmaceuticals, petrochemicals and microelectronics. A variety of experimental techniques are deployed in these facilities including photoemission and spectromicroscopy, macromolecular crystallography, low-angle scattering, dichroic absorption spectroscopy, and neutron and x-ray imaging. Applications are numerous and varied. For example, crystallography reveals the structures of viruses and proteins which are important for the development of new drugs to fight everything from flu to HIV and cancer. Penetration deep inside materials such as steel can identify stresses and strain within engineering components such as turbine blades. Tomography investigations reveal microscopic details of the 3D-structure of the brain. Observation under changing conditions can help improve process for the manufacture of plastics and foods and develop ever smaller magnetic recording materials important for data storage in computers.

The digital revolution has enabled rapid advances and opened up huge opportunities for all these research fields while at the same time bringing some significant challenges. The research community has begun to address unresolved challenges in long-term preservation and access to information by setting up repositories, some focusing on documents, some on data, others on both, with many serving specific disciplines, and devising sound policies to encourage the sharing of the data. Whilst the more general aspects of European data infrastructure are being coordinated by various initiatives and projects such as the Alliance for Permanent Access, e-IRG, ESFRI, many of which involve the PaNdata partners, the PaNdata ODI project addresses the specific, urgent, and pragmatic needs for a data infrastructure serving the Photon and Neutron science communities in Europe.

The participating facilities serve an expanding user community of well in excess of 30,000 visiting scientists each year across Europe and are major producers of scientific data. Three

new light sources became operational relatively recently (SOLEIL, DIAMOND, PETRA-III) and several other facilities are being planned, under construction or upgrade (ALBA, EU-XFEL, FERMI, ESRF, ILL, ISIS, SwissFEL). Taken together these facilities will soon produce enormous quantities of scientific data, more, for example, than is planned for the Large Hadron Collider (LHC) at CERN. This upcoming "data avalanche", a result of the increased capability of modern electronic detectors and high-throughput automated experiments, makes it essential that forces are joined to implement and deploy a framework for efficient and sustainable data management and analysis.

The facilities are in the centre of scientific activity of this community proving a focus to activities and producing the data which are the raw materials for science. The experiments in these facilities are of increasing complexity, and increasingly performed in more than one laboratory by collaborations between international research groups. The resulting raw and processed data need to be accessible over the Internet across facilities and user institutions. It should remain on-line at least until the results are published, in many cases much longer to allow re-processing and the preservation of knowledge.

Historically, the situation at many of the facilities, and in particular at the photon sources, has left data management largely up to the individual users who often literally carried data away on portable media. These media are notoriously unsuitable to guarantee the longevity and availability of precious and costly experimental data. Not only is this becoming unfeasible considering the dramatic increase in size of some of the data sets, it is also counterproductive for the scientific workflow, verifiability of the data analysis and ultimately constitutes a dramatic loss for the scientific community. Presently, access to instruments, data, software and e-infrastructure is being standardised between the facilities through the PaNdata Support Action. This will tremendously simplify the landscape for multi-disciplinary exploitation of the instruments and lay the groundwork for common implementation of data management infrastructure across these participating facilities and beyond.

Once agreement is reached on data standards for European synchrotrons and neutron sources and implemented through open networked interfaces, this will allow industry to utilise publicly available data, processing or reordering the data in such a way that it could be presented with added value to commercial market segments like, for example, life science, engineering or material science.

The potential and progress of the project will be readily disseminated to the scientific community through other relevant Integrated Infrastructure Initiatives (I3), specifically, NMI3 for neutrons and ELISA for synchrotrons and FELs. NMI3 and ELISA are each coordinated by one of the partners of PaNdata Europe. Links to other relevant types of multidisciplinary RIs, such as lasers or NMR, will be made through the I3 Network which is also coordinated by one of the partners. These will also enable rapid roll-out to other neutron and photon RIs. Cooperative knowledge exchange between PaNdata and e-infrastructure providers like EGI and PRACE will strongly benefit from the standardisation efforts and significantly enhance the research opportunities of photon and neutron user communities.

The clear benefit of an EU-funded collaborative project will be the strong incentive and timescale for initiating and completing actions. Considering the demonstrated success of collaborative ventures within the NMI3 and ELISA projects and their successful routine operation, we expect the same to evolve from this work. This project also provides an opportunity for wider collaborations between similar relevant European initiatives and will ensure integration into the wider data infrastructure supporting multi-disciplinary science. And last but not least, PaNdata ODI will stimulate discussions and possibly collaborations with North American neutron and photon laboratories which are currently lacking similar initiatives.

1.1.2 Photon and Neutron science

Photon and Neutron laboratories work naturally in close relationship. This is reflected in the physical proximity of ISIS and Diamond at the Rutherford Appleton Laboratory, the ILL and the ESRF on a joint site of the Polygon Scientific in Grenoble, the SINQ and SLS at the Paul Scherrer Institute in Villigen, and by the creation last year of the HZB laboratory combining the HMI and BESSY in Berlin. PaNdata will further reinforce such collaboration by sharing expertise and best practices in data management across both communities.

To drive the development and evaluate the benefit of the services deployed, PaNdata ODI will implement three virtual laboratories which provide case studies in the use of the shared data infrastructure. These virtual laboratories will support the following techniques:

1. structural 'joint refinement' against X-ray & neutron powder diffraction data,
2. simultaneous analysis of SAXS (Small Angle X-ray Scattering) and SANS (Small-Angle Neutron Scattering) data for large-scale molecular structures,
3. tomography such as demonstrated in the rendering of palaeontology samples.

The figures on the three following pages provide short vignettes these techniques.

1.1.1 Impact of PaNdata in Europe and beyond

Keeping track of experimental data is becoming an increasingly important part of the scientific process as the rate at which experiments can be performed and analysed is increasing. With more software tools being written to take advantage of experimental data from more than one source to deliver a more accurate portrayal of 'the material world', the ability to source this data quickly and easily becomes increasingly important. Furthermore the increasingly global nature of scientific collaborations requires researchers from different organisations to seamlessly work with data from more than one source. These complex interactions place increasing taxing demands on researchers to demonstrate the provenance of data and analysis applied to it.

The partners in this proposal are not only providers of 'hardware-based' experimental facilities for users, but also of associated software tools, algorithms, computational resources etc. As such, they are ideally placed to impact markedly upon the scientific method by enabling the provision of facility-derived data technology not only to their own users but also to the wider scientific community.

Sitting at the heart of this vision is a series of catalogues, which allow users to perform cross-facility, cross-discipline interaction with experimental and derived data, with near real-time access to the data. Associated with these data catalogues, and highly cross-referenced with them are further catalogues of users, publications, and data analysis software. Together, these ensure controlled access to files and the ability to track dependencies from data to publication and vice-versa. Taken together, these catalogues and their associated linking technologies, point the way towards a major change the way in which users will interact with their data before, during, and after a facility experiment. They will also through wider accessibility and long-term availability of data and through use of common languages and tools, encourage and support new interdisciplinary research.

This project will bring together the information infrastructures of major research facilities. This is a significant step along the road to a fully integrated, pan-European, information infrastructure supporting the scientific process. This step is not only important because of its technological benefits, but is also essential because on the sociological side it will bring along with it the very significant scientific community which uses these Research Infrastructures (RIs).

Virtual Laboratories Task 1:
Structural joint refinement against X-ray and neutron powder diffraction data

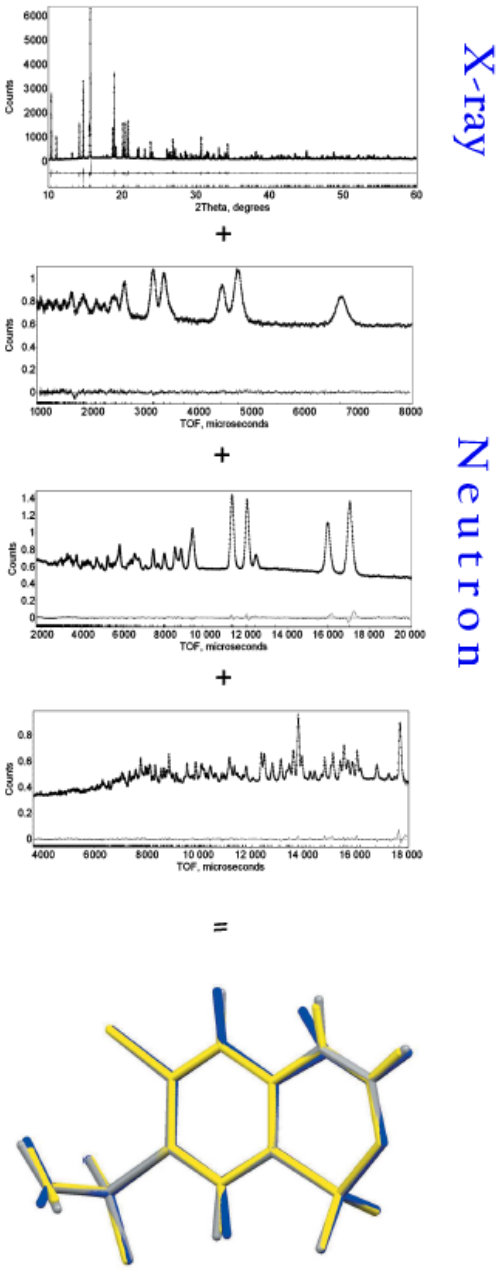


Figure 1: *XRPD data collected on ID31 at the ESRF is combined with multibank neutron powder data from the GEM diffractometer at ISIS to give a refined structure (grey) for fully protonated chlorothiazide. The single crystal X-ray structure is shown in yellow.*

X-rays and neutrons provide highly complementary information in the context of crystal structure determination and refinement, as a result of the significant differences between X-ray scattering factors and neutron scattering lengths for contributing atoms. The archetypal example is that of the hydrogen atom, whose nuclear position can be accurately determined by neutron scattering but not by X-ray scattering. Combining X-ray (for heavier atoms) and neutron (for hydrogen) scattering data (suitably collected) delivers a level of accuracy and precision in a structural refinement that exceeds that obtainable from either single source taken in isolation.

Such combined usage will be greatly facilitated by the use of federated metadata catalogues that allow datasets for particular compounds to be located, even when they have been collected at different facilities. Careful use of sample descriptors (using suitable ontologies where appropriate) will be a key component of successful searching, as will the ability to reference reduced data as well as raw data. In the field of crystallography, reduced data is generally in a simple format, such as *xye* files for powder data; such files can be retrieved and fed directly into standard structure refinement packages such as GSAS. This concept is easily extended to the analogous single-crystal situation, where reduced data in simple formats (e.g. SHELX HKL) gleaned from disparate sources can be combined in a single refinement.

Virtual Laboratories Task 2:
Simultaneous analysis of SAXS and SANS data for large-scale molecular structures

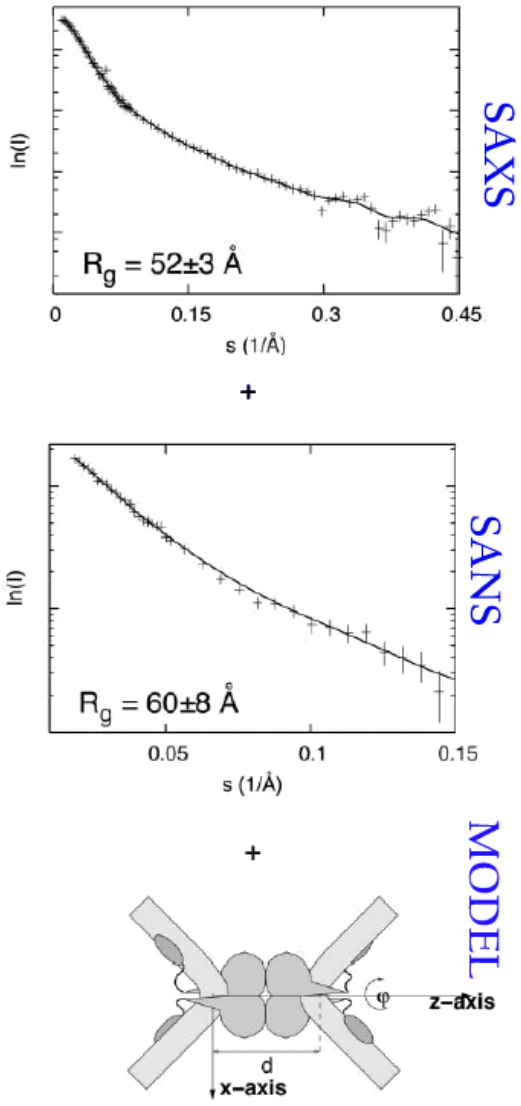


Figure 2*: SAXS data (BL 2.1, Daresbury SRS) and SANS Data (D11, ILL) have been modelled to give the solution structure of the NM36 X synapse. In the proposed work package, data collected on I22 at Diamond and SANS2D at ISIS will form the core of the study.*

Small-angle scattering is an extremely valuable technique for probing the nanoscale and mesoscale (as opposed to the atomic scale) structure of materials and, in particular, soft condensed matter. For example, it can be used to return size, shape and ordering information on systems as diverse as macromolecules, polymers, liquid crystals and vesicles.

Critically, such small-angle scattering approaches can be used to study molecules and assemblies *in solution* (as opposed to in the crystalline state) and as such, the behaviour of systems can be studied as a function of exposure to a wide range of solution conditions such as pH and salt concentration. The use of synchrotron X-rays helps to compensate for weak scattering from dilute solutions, though there is always a risk of radiation damage. Neutrons scatter more weakly but with no risk of radiation damage and they also allow use of contrast matching techniques. SANS and SAXS are thus highly complementary and are increasingly likely to be used in combination in detailed studies of nano- and mesoscale structures.

The ability to locate, download and analyse SAXS/SANS data collected from large-scale structures will not only encourage and tremendously facilitate such combined analysis but will also encourage proposals for future experiments, by allowing users to see what has been / can be achieved using currently available data.

Virtual Laboratories Task 3: Tomography data repository for palaeontology samples

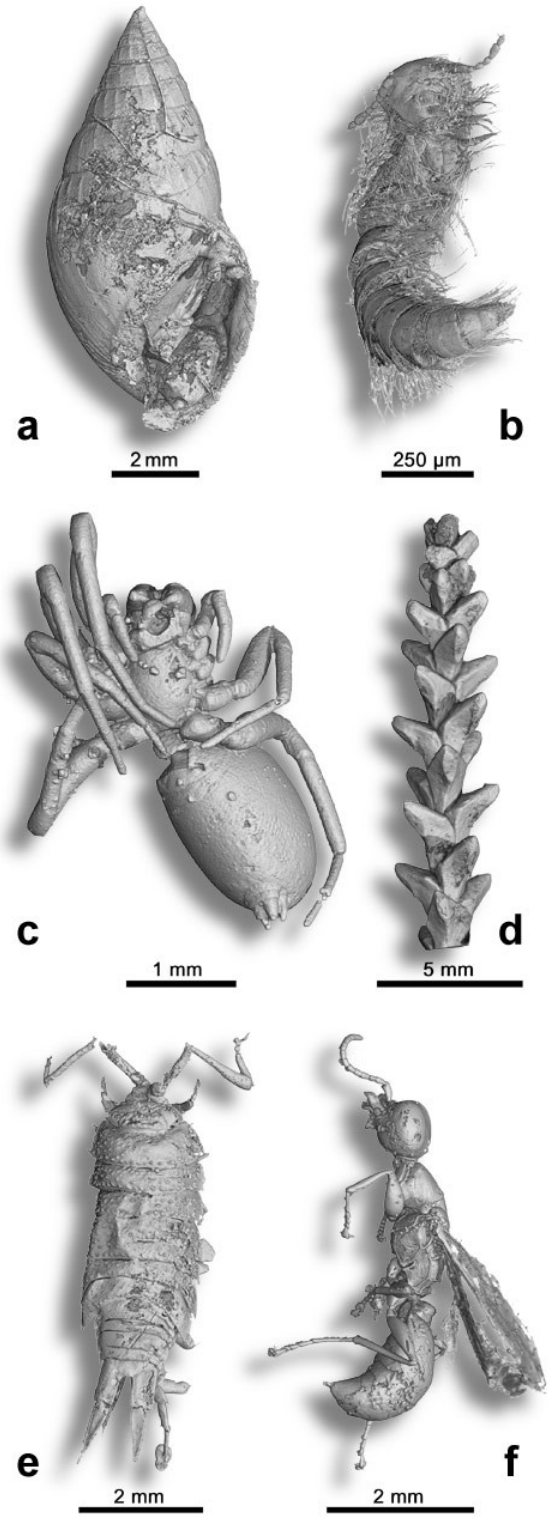


Figure 3: Examples of virtual 3D extraction of organisms embedded in opaque amber: a) Gastropod Ellobiidae; b) Myriapod Polyxenidae; c) Arachnid; d) Conifer branch (Glenrosa); e) Isopod crustacean Ligia; f) Insect hymenopteran Falciformicidae. Credits: M. Lak, P. Tafforeau, D. Néraudeau (ESRF Grenoble and UMR CNRS 6118 Rennes).

Amber has always been a rich source of fossil evidence. X-rays now make it possible for palaeontologists to study opaque amber, previously inaccessible using classical microscopy techniques. Scientists from the University of Rennes (France) and the ESRF found 356 animal inclusions, dating from 100 million years ago, in two kilograms of opaque amber from mid-Cretaceous sites of Charentes (France). In a second study, synchrotron X-rays were used to determine the 3D structure of feathers found in translucent amber, to complement the information already known about the feathers. The feather fragments are unique because they may have belonged to a feathered dinosaur featuring feathers in an intermediate stage of evolution to those of modern birds.

Palaeontology is a new research field using X-rays for non-destructive examination of samples. Samples measured at synchrotrons should be deposited in a database and can be made easily publicly accessible after the results have been published. Depending on the kind of sample, the data for each sample represents between 2 and 100 GB. The data will have to be properly annotated with the technical acquisition parameters, the details about the sample itself as well as the processing information. Finally, it needs to be linked to the relevant publication or contain at least the reference to the publication. A palaeontology database would be supplied with several TB of data per year. Secure authentication and access for data deposition as well as secure archiving of the data are issues which must be addressed.

The potential and progress of the project will be readily disseminated to the scientific community through the relevant Integrated Infrastructure Initiatives (I3), specifically, NMI3 for neutrons which is coordinated by one of the partners, and the ELISA project for synchrotrons which is also coordinated by one of the partners. These will also enable rapid roll-out to other neutron and photon RIs.

The clear benefit of an EU-funded collaborative project will be the strong incentive and timescale for initiating and completing actions. EU funding will allow help remove the usual barriers of choosing and adopting standards between partners, inherent to all software collaborations. Considering the demonstrated success of collaborative projects within the NMI3 and ELISA projects and their successful routine operation, we expect the same to evolve from this project. This project also provides an opportunity for wider collaborations between similar relevant European initiatives and will ensure integration into the wider data infrastructure supporting multi-disciplinary science. And last but not least, PaNdata will stimulate discussions and possibly collaborations with North American neutron and photon laboratories where currently no similar initiative exists.

1.1.2 Consortium

PaNdata brings together the data infrastructure providers from some of the largest multidisciplinary RIs in Europe to develop common technology and practices and evolve towards a single user experience for their communities. These RIs already now share much in common. They operate hundreds of instruments for experiments which provide a wide variety of information from the scale of atoms to the scale of ants, in materials ranging from proteins to turbine blades. They are used by well in excess of ten thousand scientists each year, with overlapping constituencies of users, for thousands of experiments and have demand far beyond their capacity. The two RIs based in Grenoble are international organisations whilst the others are primarily national funded, though many have significant international use (e.g. more than half of the PSI and ELETTRA users are international). They are all world class. These similarities provide a common basis and understanding that will enable rapid progress. There are also some critical historical differences between the RIs, in terms of technologies used or policies applied, which will ensure that the technology and practices developed in this project will be generic and thus applicable to a wider range of facilities in the future. All partners will actively contribute in defining the work of the consortium and in deploying and serving the outcome to the user community.

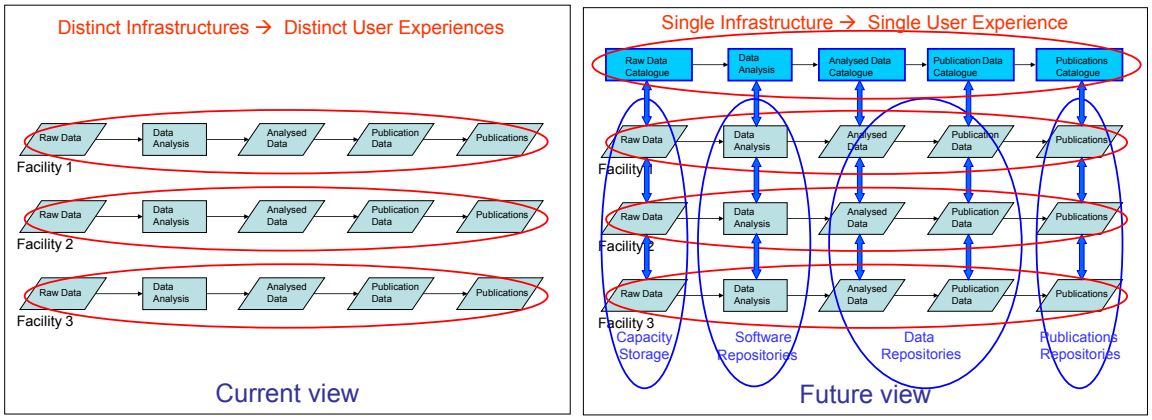
The UK RIs have a close working relationship with a large e-Science department which is highly active in providing infrastructural software technology for scientific research in the UK and Europe. The involvement of the STFC e-Science centre ensures awareness and compatibility with related activities in environmental sciences, particle physics, astronomy and social science and thereby prepares the ground for integration into a wider European data infrastructure. STFC e-Science also coordinates the UK activities in EGEE and EGI ensuring that relevant infrastructure for authentication and data access can be leveraged.

The consortium is particularly well balanced, being diverse enough to ensure that results have broad applicability, yet focused enough to deliver effective results quickly and within a reasonable budget.

1.1.3 Conceptual design

Our vision is to standardise and integrate our research infrastructures in order to establish a common and traceable pipeline for the scientific process from scientists, through facilities to publications. At the heart of the vision is a series of federated catalogues which allow scientists to perform cross-facility, cross-discipline interaction with experimental and derived data, with near real-time access to the data. This will also deliver a common data management experience for scientists using the participating infrastructures particularly fostering the multi-disciplinary exploitation of the complementary experiments provided by neutron and photon sources.

Building on the unification of data management policies and adoption of common data standards developed in the PaNdata Support action, this project will develop and deploy the common technologies which will realise the benefits of standardisation. The aim is illustrated in the following diagram (Fig. 1). According to the current view, each facility handles separately the full data management sequence, from generation of raw data to publication of results. In the future view, a single user experience is enabled through the use of shared technologies at the different facilities. It is clear that the common data management scheme will offer many synergies and allow completely novel possibilities.



PaNdata Vision: Current and future views of data pipeline at facilities

Such a unique infrastructure will enhance all research done in this community, by making data accessible, preserving the data, allowing experiments to be carried out jointly in several laboratories and by providing powerful tools for scientists to remotely interact with the data.

The data pipeline

The architecture follows the data pipeline from data creation through to publication of analysed results which feedback into new research proposals. The design is based on a layered approach, with well defined application programming interfaces (API's) providing communication between layers. This layered approach allows each site the choice of different implementations for the same layer to take into account local differences between sites and to optimise overall performance.

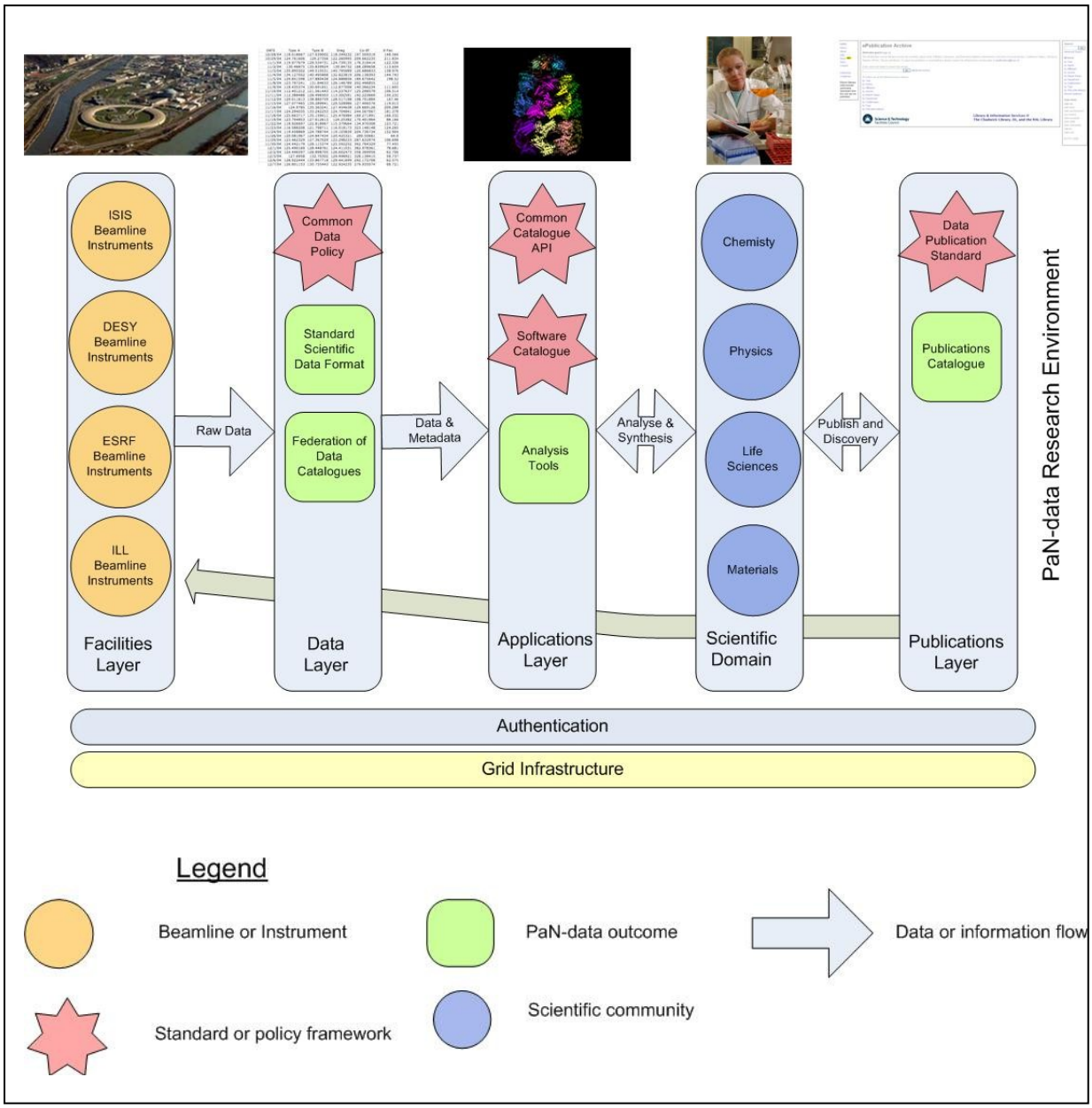
The PaNdata architecture identifies the following components:

– *Facilities Layer* – unifies the different instruments at the different facilities. This is the layer the users interact with directly while performing the experiments.

– *Data Layer* – provides a standardised environment to store and archive data. Building

upon data policies and standard data format, this layer provides the basis for the analysis framework.

- *Applications Layer* – defines the analysis framework for archived data. Data and information queries are enabled through federated catalogues and common API, which provide the basis to analyse the data utilizing the software framework implementation based on a central software catalogue.
- *Scientific Domain Layer* – unifies the results of the data analysis across disciplines.
- *Publications Layer* – completes the lifecycle and provides a uniform, public access to the results of experiments and data analysis building upon a federated publication database.



The layered structure and building blocks of PaNdata's infrastructure

These "vertical" layers are supported by common "horizontal" layers which enable transparent interaction across the activities.

- *Authentication Layer* – this layer will identify, authenticate and authorise users to access (or not) the research infrastructure and provides the bases for the layer building on top of it.
- *Grid Infrastructure Layer* – the Grid layer will be provided by other initiatives outside this proposal such as EGI supported through the NGI's and SSC's and will not form part of the work of this project, although it will build on this infrastructure as required.

1.1.4 Goals and Objectives

Neutron and photon RIs are major creators of scientific data. These data, leading on to scientific publications and knowledge, are one of their major outputs. The neutron and photon RIs in Europe are truly world class and frequently world leading. They are a core component of the European Research Area and Europe should demand that the data they produce are maximally exploited.

The overarching aim of this project is to enable new and better science by establishing common practices, services and technologies for the management of data across the participating RIs and to promote these benefits to other similar establishments.

Goals

The first goal of the project, implemented though the Networking activities, is to share existing knowledge between the partners, the users of their facilities and the wider scientific community. Building on the similarity of purpose and commonality of practice across the participating facilities, there are many areas of practice with regards to data handling where the formulation of a cohesive framework would be beneficial to the partners, similar organisations, and the scientists using them.

The second goal of the project, implemented though the Service Activities, is to deploy and operate a common set of services for catalogued access to scientific data which provide provenance information and managed preservation and which, in turn enable the development of new services across raw, analysed or published data which will be the real scientific merit. Given the fact, that there is a significant overlap of users and scientific applications, such commonality is high on the priority list for facility users.

The third goal of the project, implemented though the Joint Research Activities, is to provide a managed package of open source software available to the partners and to other facilities that will support the establishment of repositories of scientific software built upon new and existing components. Given constraints on resourcing available, not all the partners will contribute to all the areas of work, although all will benefit from all the outcomes.

Objectives

Over past 3 years, the PaNdata collaboration, both independently and through the support of the FP7 programme in the PaNdata Support Action project and other joint projects, have undertaken a programme of standardisation and strategic planning which has detailed a programme of activities working towards the construction and operation of a shared data infrastructure for Neutron and Photon laboratories across Europe. The proposed project will build on this work to address this roadmap, exploit synergy, and deliver common open data infrastructure shared across the participating facilities.

The objectives of the proposed project are detailed below.

Objective 1 – Collaboration

To establish an effective and efficient collaboration between the partners delivering added value to each participant through shared research, service and networking activities and to integrate this collaboration with related infrastructure initiatives beyond the project.

Outcomes

Specifically we will:

1. undertake joint networking, research, and service activities leading to collaborative specification, development and operation of the developments and services,
2. agree on appropriate common definitions and policies required to achieve the goals of the project,
3. monitor progress of these joint activities and put in place appropriate corrective actions if this progress falls short of that required to deliver the project,
4. prepare and deliver the outputs and deliverables defined in this project plan,
5. ensure effective communication of project outputs to facility user communities, partner RIs and more general (e-)infrastructure developments,
6. engage with related e-infrastructure and data integration developments outside the project, in particular across Europe, with a view to the longer term integration of this work into the broader integrated infrastructure required to support European Research in the coming decade,
7. contribute to the development of the broader infrastructure through participation in relevant integration, planning and standardization activities required to achieve the eIRG vision of an integrated European e-Infrastructure.

Objective 2 – Users

To deploy, operate and evaluate a system for pan-European user identification across the participating facilities and implement common processes for the joint maintenance of that system.

Outcomes

Specifically, we will:

1. develop a generic infrastructure to support the interoperation of facility user databases enabling unique identification of users and supporting authentication across the facilities and with other similar infrastructures in the wider context,
2. deploy this infrastructure to establish a single catalogue of users across the partners,
3. provide a user login service based upon this generic framework which will enable users single sign on to partners' systems,
4. evaluate this service from the perspective of facility users,
5. manage jointly the evolution of this software and the services based upon it,
6. promote the integration of this technology and services based upon with similar systems beyond the project.

Objective 3 – Data

To deploy, operate and evaluate a generic catalogue of scientific data across the participating facilities and promote its integration with other catalogues beyond the project.

Outcomes

Specifically, we will:

1. develop the generic software infrastructure to support the interoperation of facility data catalogues,
2. deploy this software to establish a federated catalogue of data across the partners,
3. provide data services based upon this generic framework which will enable users to deposit, search, visualise, and analyse data across the partners' data repositories,
4. evaluate this service from the perspective of facility users,
5. manage jointly the evolution of this software and the services based upon it,
6. promote the take up of this technology and the services based upon it beyond the project.

Objective 4 – Provenance

To research and develop a conceptual framework, defined as a metadata model, which can record the analysis process, and to provide a software infrastructure which implements that model to record analysis steps hence enabling the tracing of the derivation of analysed data outputs.

Outcomes

Specifically, we will:

1. develop a framework which allows logging of processes undertaken by scientists in data analysis,
2. develop ontologies for specific disciplines and specific techniques to instantiate the framework

Objective 5 – Preservation

To add to the PaNdata infrastructure extra capabilities oriented towards long-term preservation and to integrate these within selected virtual laboratories of the project to demonstrate benefits. These capabilities should, as for the developments in the provenance JRA, be integrated into the normal scientific lifecycle as far as possible. The conceptual foundations will be the OAIS standard and the NeXus file format.

Outcomes

Specifically, we will:

1. apply and adapt the OAIS standard to the data holdings of the partners' facilities,
2. define a common schema for preservation-oriented metadata to support the application of the OAIS standard,
3. track the development of the NeXus International Standard format with respect to preservation requirements,

4. develop tools and techniques to integrate the capture and propagation of the metadata into the catalogues of users, data, software and publications,
5. develop methods for recording analysis software as part of the preservation metadata,
6. evaluate the effectiveness of the above in enabling reuse and long-term preservation,
7. manage jointly the evolution of the formats and schemas and the software tools based on them,
8. engage with international standardisation efforts to promote the take-up of these standards and services based on them beyond the life of the project.

Objective 6 – Scalability

To develop a scalable data processing framework, combining parallel filesystems with a parallelized standard data formats (pNexus pHDF5) to permit applications to make most efficient use of dedicated multi-core environments and to permit simultaneous ingest of data from various sources, while maintaining the possibility for real-time data processing.

Outcomes

Specifically, we will:

1. develop a pNexus/pHDF5 API
2. implement this on specific parallel filesystems
3. demonstrate its use for selected applications (tomography, crystallography)

Objective 8 – Demonstration

To deploy and operate the services and technology developed in the project in virtual laboratories for three specific techniques providing a set of integrated end-to-end data services.

Outcomes

Specifically, we will:

1. deploy virtual laboratories for three example techniques:
    a. Structural 'joint refinement' against X-ray & neutron powder diffraction data
    b. Simultaneous analysis of SAXS and SANS data for large scale structures
    c. Tomography data exemplified by paleontological samples
2. evaluate this service from the perspective of facility users,
3. manage jointly the evolution of this technology and the services based upon it,
4. promote the take up of this technology and the services based upon it beyond the project.

1.1.5 Outline programme of work.

The programme of work is broken down into 8 work packages which together cover the spectrum of activities required to enable the conceptual design and objectives described above. Some work packages are technologically focused concentrating on the research and development required to bring new technologies up to production quality. Some are concerned with the deployment and operation of that technology, whilst others address the coordination aspects required to effectively put the new technology into practise.
The work packages address the following topics:

Networking Activities

1. **Management** and related activities
2. **Engagement** with other initiatives and **dissemination** of project outcomes

Service Activities

3. Deployment, operation and evaluation of a common **AAA service for users**
4. Deployment, operation and evaluation of a common metadata **service catalogue**
5. Deployment, operation and evaluation of a common virtual laboratories serving specific case study techniques.

Joint Research Activities

6. Research and development of shared technology for **Provenance**
7. Research and development of shared technology for **Preservation**
8. Research and development of shared technology for **Scalability** of data transfer

1.1.6 Relation to topics addressed by the call

"Increase of the scale of federation and interoperation of data infrastructures,..."
The project will undertake the research, development, deployment and operation of a common scientific data infrastructure across the participating facilities. In doing this, it will foster the transition from local and national solutions addressing the immediate demands of the individual neutron and photon facilities, to a harmonised approach across the participants and other European research infrastructure providers. By providing a coordinated deployment of a common set of data related services across these facilities, it will contribute significantly to the deployment of a European scientific data infrastructure and towards the development of common infrastructure with similar initiatives on other continents.

"... better exploitation of synergies with the underlying e-Infrastructures, reduction of costs, increase of the user base ..."

The project will bring together the expertise of some world leading research facilities and so promote best practice in data management between the participating facilities and, by example, encourage the emergence of this best practice into the wider community. Besides enhancing the efficiency of the data individual fields, standardisation will be an enormous source for synergy by knowledge exchange between teams from the wide research fields active at the facilities. This will stimulate the emergence of new working methods and engender the development of a new research environment. It will therefore add value to the outputs of the facilities both in terms of scientific performance and extent of access. As an example, a common users catalogue and AAA service will for the first time allow a systematic study of efficiency and dedicated optimisation of the impact of the e-infrastructure services developed.

"... bridging across disciplines, enabling of cross-fertilisation of scientific results and favouring of innovation."
By providing easy-to-use, controlled access to data holdings of the partner facilities, PaNdata ODI will provide a unique distributed scientific resource which will support the emergence of new working methods. Data analysis is a key link in the chain of events that transforms original ideas into conclusive scientific output. By providing traceability of data provenance through the analysis stages PaNdata ODI will lay an important step in the development of a software infrastructure which will ultimately enable the most appropriate software to be used independently of where the data is collected and therefore accelerate the deployment and use of new data analysis methods which will open doors to new science across the facilities and the user community. Because of the important role the partners play in European Photon and Neutron based science, this work will form a significant contribution to the development of a European strategy for scientific software.

"... The removal of important obstacles concerning the open access to scientific information and data, ..."

The project will promote a common user experience across the participating facilities. It will lower the learning threshold for initial use of these facilities and the transfer of expertise between them. In this way it will lead towards making the infrastructure layer transparent by hiding the complexity and distribution of the underlying systems. It will therefore both enable researchers focused on one domain to fully exploit their scientific expertise rather than "battling" the technology which is essential to their productivity, whilst also enabling cross-disciplinary scientific activities by facilitating access to research across fields.

"...as well as the improvement of preparedness to face the data "tsunami" of the next decade."

PaNdata ODI will investigate the development of scalable data flow frameworks that support the accumulation of data from several detectors through data formats which build on parallel filesystems and protocols. These developments can serve as a proof of principle to guide developments for x-ray free electron lasers and other new facilities coming up in future years in which the partners are involved.

"Progress towards the vision of open and participatory data-intensive science."

The infrastructure developed will be ultimately inclusive, readily integrating related national and international facilities, as well as collaborative, looking to exploit synergies with other data infrastructures relevant to the research communities served. It will also engender more intense collaborations between the research infrastructure providers and the researchers in their virtual research communities, to share and exploit the collective power of the European landscape of Photon and Neutron facilities.

PaNdata partners are simultaneously participants in ESFRI projects such as ESRFUP, ILL 20/20 and IRUVX and there are already intense discussions going on to efficiently synchronise these activities. A similar situation occurs in the neutron (NMI3) and photon (ELISA) I3 access programs. Here cooperation between projects is foreseen in the form of cross exchange of delegates at the respective plenary meetings. By establishing collaboration between the participating organisations, this Support Action will provide a unique platform from which to disseminate the work of this and other projects in the e-infrastructure programme.

Traditionally, there are active collaborations in many fields of the European neutron and photon community with overseas partner facilities in the US and Canada. It will be natural to share with our colleagues the results of novel e-infrastructure developments. This will encourage efficient e-infrastructure cooperation on the global scale including roadmapping. As there is no reason for restricting a successful realisation of this e-infrastructure project to the European scale, the cooperative development of technologies and services established within this consortium will provide an important step towards similar cooperation at a global scale.

1.2 Progress beyond the State of the Art

This section describes the current status of data/information management at each of the participating facilities and the advancements that the project is expected to bring through the underpinning technology which we will build and deploy in the project.

1.2.1 State of the Art at the participating organisations

State of the Art at STFC/ISIS

Experiments on instruments at ISIS (http://www.isis.rl.ac.uk) are controlled by individual instrument computers, closely coupled to data acquisition electronics (DAE) and the main neutron beam control. Beyond the initial production of RAW neutron data, this control breaks down into a series of more discrete steps.

- Experiments generate RAW (ISIS specific) files, which are copied to intermediate (central archive) and long term (ATLAS tape robot) data stores for preservation.
- Annotation of the RAW data is limited; search / retrieve of stored data is largely achieved by browsing or by use of specific experiment run numbers.
- Access to RAW data is controlled at the instrument level.
- Reduction of RAW files, analysis of intermediate data to generate results and publication of those results is a process that is largely decoupled from the handling of the RAW data
- Valuable connections in the chain between experiment and publication are not preserved.

Future data management at ISIS will focus on the implementation of a loosely coupled set of self-contained components that have well-defined and standardised interfaces; this allows for a far more complex / flexible set of interactions between components

- The ICAT metadata catalogue[1] sits at the heart of this strategy, controlling access to files and metadata, implementing a clear data policy and using SSO for authentication.
- Communication between components is achieved using web services and ODBC.
- User space is now much more closely aligned with facility space.
- Component development is simplified and can be distributed between different groups
- The RAW file format will be replaced by the Nexus format.
- ICAT allows linking of all types of data, from beamline counts through to publication data

ISIS ICAT will be one of many facility ICATs that can be searched simultaneously via a WWW-based data portal "TopCat".

[1] http://code.google.com/p/icatproject/

State of the Art at ESRF

The European Synchrotron Radiation Facility (http://www.esrf.eu) is a third generation synchrotron light source, jointly funded by 19 European countries. It operates 40 experimental stations in parallel, serving over 3500 scientific users per year. At the ESRF, physicists work side-by-side with chemists, materials scientists, biologists etc., and industrial applications are growing, notably in the fields of pharmaceuticals, petrochemicals and microelectronics. It is the largest and most diversified laboratory in Europe for X-ray science, and plays a central role in Europe for synchrotron radiation. ESRF provides the computing infrastructure to record and store raw data over a short period of time and also provides access to computing clusters and appropriate software to analyse the data. The ESRF will witness a dramatic increase in data production due to new detectors, novel experimental methods, and a more efficient use of the experimental stations. The "Upgrade Programme", a science and technology programme to push a significant part of the ESRF beamlines to unprecedented performances, will further increase the data production from currently 1.5 TB/day by possibly three orders of magnitude in ten years from now. The ESRF is currently reviewing its data management scheme in view of possibly implementing long term storage of curated data for in-house research projects. The long-term preservation and access to scientific data will constitute a major challenge for the photon and neutron science community. Data policies need to be addressed community wide and the necessary tools can only be developed on a European scale.

The ESRF has a long track record of successful international collaborations in many different fields of science and technology (SPINE, BIOXHIT, eDNA, X-TIP, SAXIER, TOTALCRYST, etc.). Three international projects are of direct relevance to PaNdata – the international TANGO control system collaboration, ISPyB, and SMIS:

The TANGO control system was initially developed for the control of the accelerator complex and the beamlines at ESRF and has been adopted by SOLEIL, ELETTRA, ALBA, and DESY. The TANGO collaboration does not rely on external funding. It shows that five of the PaNdata partners are already working together in software developments of common interest.

ISPyB is part of the European funded project BIOXHIT for managing protein crystallography experiments. In its current state, it manages the experiment metadata and data curation for protein crystallography. PaNdata intends to go much further because it addresses data from all experiments. We will exchange information with the ISPyB project to make sure there is no duplication of effort.

The SMIS project is the ESRF's database for handling users and experiments; it does not yet handle data or metadata, but the scheme envisaged here will allow it to be fully integrated into a larger data management scheme.

The ESRF will support the proposed project beyond the requested funding from FP7 in the following ways:

- The hardware infrastructure for storing, analysing, and archiving data, as well as all the hardware required for participating in the PaNdata photon and neutron GRID initiative will be sourced from the ESRF annual budgets.
- Modifications or adaptations of the ISPyB and SMIS, as well as other software packages will also be sourced from the operations budget of the ESRF.

State of the Art at ILL

The ILL (http://www.ill.eu) has a fully-functional computing environment that covers all aspects of experiment and data management; most of the tools have been running for many years and continue to evolve, but they are not shared with any other RI. The main points of the current system are briefly described below.

All neutron data since the start of the ILL is stored. Data collected since 1995 is easily available using Internet Data Access (IDA, see below) All data is stored in ILL ASCII format. The two exceptions are the new instruments BRISP and IN5, which generate data sets that are too large to store, but above all, too slow to read. BRISP is the first ILL instrument using the NeXus format. The Instrument Control Service has developed a module that generates NeXus files from its internal format: this module is valid for all instruments, allowing all ILL data to be converted to NeXus, once the contents have been defined. Internally, data can be accessed directly on the central repository. Most users take a copy of their data when they leave but they can log-in from their home labs and retrieve data by direct methods (SFTP, SCP …) or using IDA (barns.ill.fr), which has run for almost 10 years and is reasonably well used. A new catalogue and the interconnection of the catalogue of the different European facilities will be of great help for our users.

Since the beginning of 2010, ICAT the new data catalogue has been rollout with access restricted to the ILL staff scientists. Once the ILL Data policy based on PaNdata work will be release, ICAT will replace IDA and will be made accessible to our users.

The Scientific Coordination Office (SCO) has a data base of users and the "ILL Visitors Club" is a user portal which constitutes a web-based interface to the SCO Oracle database.

All administrative tools for ILL users are grouped together and directly accessible on the web in the Visitors Club. On entering a personal and unique ID, a user's personal details are automatically recalled and they can access directly all the available information which concerns them. They can also update their personal information.

The data base (and the information stored in it) is shared by different services at the ILL (site entrance, welcome hostesses, health physics, reactor guardians, etc.) through different web-interfaces and search programs adapted to their needs.

The ILL Visitors Club includes the electronic proposal and experimental reports submission procedures and makes available additional services on the web, such as acknowledgement letters, subcommittee electronic peer review, subcommittee results, invitation letters, instrument schedules, user satisfaction forms and so on.

Utilisation of the technologies envisaged in this proposal will of course impact very favourably upon the compatibility of ILL data and information with that of the other partner facilities. Of particular import will be adoption of NeXus format across the facilities, as this will enable major data analysis programs (such as the SANS-suite (Fortran), Mfit/Mview (Matlab) and LAMP (IDL)) to be brought to bear of more diverse data sources. Existing couplings between ILL databases will be strengthened (e.g. proposal through to publication) and exposure of ILL data and resources will be significantly improved

State of the Art at Diamond

Diamond Light Source (http://www.diamond.ac.uk/) is a new 3rd generation synchrotron light source. It is the largest scientific facility to be funded in the UK for over 40 years, and became operational in January 2007. Diamond is in the advantageous position of being able to profit from the hard won experience of other facilities while actively commissioning many X-ray beamlines during the period covered by the proposal. Currently there are 11 user scheduled beamlines available with 4 new beamlines being commissioned each year and the active user population is growing rapidly and will soon exceed 1000 users drawn from the UK, the rest of Europe and indeed the rest of the world.

The state of the art:

- The same underlying JAVA based Generic Data Acquisition (GDA) system is used globally but has been configured for the specific scientific and user needs of each beamline.
- The use of Java enables direct integration with many software packages already available.
- The low level control system is the widely used EPICS system which provides a stable and reliable means for hardware control.
- Diamond has worked closely with ISIS, our Central Laser Facility, e-Science and the central site services to implement a cross site user authentication database.
- Diamond has collaborated with the ESRF and ISIS to implement Web based user administration (DUODESK) and proposal submission (DUO) applications.
- The DUODESK application is integrated with most aspects of user operation ranging from accommodation and subsistence through to system authentication, authorization and metadata retrieval.
- We are currently working with e-Science and ISIS to provide an initial externally available data storage repository based on the Storage Repository Broker (SRB) with ICAT database. User authentication is enabled by the cross site wide user authentication database.

State of the Art at PSI

PSI (http://www.psi.ch) is hosting three large user facilities, SINQ – the Swiss Spallation Neutron Source, SμS – the Swiss Muon Source, and SLS – the Swiss Light Source. In addition, PSI is currently embarking on the SwissFEL project for a fourth large user facility for delivering hard X-rays. Parliament decision on the proposal is expected for 2011.

The current data acquisition and data storage environment is heterogeneous: various machine and beamline operational parameters are provided by the facilities but there is no standard for recording metadata.

SINQ uses the in house program SICS for data acquisition. Most SINQ instruments already store their raw data in the NeXus format. All SINQ data files ever measured are held on an AFS file system and are visible to everyone. Most files are indexed into a database searchable via a WWW-interface. The SμS facility uses the MIDAS software for data acquisition. Data files are stored in a home grown format; however in the long term all SμS data files will be written in the NeXus format. All data ever measured is also made public on an AFS file system. SμS and SINQ data analysis software is accessible remotely through a special computer outside of the PSI firewall. Data acquisition at SLS is based on the EPICS system. Data measured at SLS is stored on central storage for two months only. Users are supposed to take their data home on portable storage devices. There is only very limited support for data analysis at SLS.

Since about 10 years user management at PSI is handled via the on-site developed Digital User Office (DUO). This tool covers all aspects of a proposal system starting from proposal submission to automatically providing access for the users to the doors of the beamline hutches. First developed for the Swiss Light Source SLS, it includes now also the neutron spallation source SINQ. In the meantime, most European sources are running for their user management copies of DUO. There is, however, no exchange of information between the different DUO versions.

There is an increasing tendency at photon and neutron facilities that scientific questions cannot be answered by single experiments at single facilities but that rather results from different facilities (e.g. SINQ and SLS at PSI or SLS and ESRF) have to be combined. Furthermore, because of the large overbooking of the available facilities, users will use beamtime all over Europe wherever it is available so that different parts of an experimental project may be performed at different facilities. The current heterogeneous IT environment puts an unnecessary overhead on these experiments and unnecessary resources have to be invested for converting experimental information to different standards. Therefore, PSI is very much interested in an EU-wide data format which is essential for combining data from different experiments at PSI and other European facilities. In addition, a standard data format is prerequisite for archiving of experimental data.

Furthermore, it will be increasingly complicated to transfer the large datasets produced by the pixel detectors – especially at imaging-type beamlines – to the user home institutions. This will increase the demand for remote data analysis at the large facilities. These trends clearly ask for an efficient EU-wide user management, data file exchange and access system.

PSI sees the contribution of PaNdata mainly in the development and implementation of new tools and in initial service, whereas hardware infrastructure and operational resources for storing and analyzing data for internal and external users will be provided by the PSI budget.

State of the Art at DESY

DESY (http://www.desy.de) has a long history in High Energy Physics (HEP) and Synchrotron radiation. DESY runs a Tier-1 centre for the LHC project (might even act as a Tier-0 in the future) and has proven expertise in the management and storage of very large data volumes. DESY jointly provides the major software framework (dCache) for large scale and secure data storage. However, the photon science community has substantially different demands than the HEP community. Data access patterns and analysis frameworks pose rather different constraints on data management and storage and the wide spectrum of experiments usually result in a wide spectrum of heterogeneous data formats.

While HEP remains an important pillar at DESY, the main focus has clearly shifted towards photon science. DESY is nowadays operating two dedicated synchrotron sources (Doris and Petra III) as well as a VUV free electron laser (FLASH). Although Petra III, the most brilliant synchrotron source worldwide, became operational only very recently, an extension of Petra III to host additional instruments is already in planning phase. The construction of the European XFEL (www.xfel.eu) is progressing well and construction of a second FLASH facility will start soon, accompanied by the foundation of a Center for Free Electron Lasers (CFEL) as well as a Center for Structural and System Biology (CSSB). In parallel, detector development is rapidly progressing, which will allow to obtain diffraction images at a sub-millisecond timescale to cope with the unique time structure of the European XFEL laser light.

These developments will boost data rates tremendously. From Petra III and FLASH we expect data volumes in the order of a Petabyte per year. The European XFEL will be capable to collect data at a rate of 200 GB per second, extending data rates by at least another order of magnitude - first experiments of CFEL at the LCLS proved the capabilities of X-FELs to generate tremendous data rates. Apart from the mere data volumes, the number of experiments performed in parallel and the number of files to cope with requires a sophisticated data management scheme. The data policy outlined by the PaNdata Europe Strategic Working group (PaNdata Europe) provides the first, most important step towards a sustainable data infrastructure. However, implementation of the data policy and standardized data formats, collection of meta-data integrated into an ontologic description of an experiment will soon become indispensable. As a first step, DESY as the lead partner of the PNI-HDRI project (www.pni-hdri.de) of the Helmholtz-Society aims to implement a generic beamline and instrument description based on the Nexus API, which can provide a suitable basis to build experiment and facility ontologies.

DESY has decided to implement an HDF5/Nexus based standard data format for all instruments, which will greatly facilitate data storage, access, retrieval and exchange between users and facilities. Since HDF5 has also been proposed by the EC as **the** standard for binary digital objects, it promises to be a sustainable choice. Additionally, HDF5/Nexus exhibits great innovative potential. The data challenge posed by the European XFEL for example can best be met by truly parallel filestreams. pHDF5 is an implementation supporting an essentially arbitrary number of parallel data streams, provided a suitable infrastructure like parallel filesystems and MPI-IO is available. DESY has already substantial know how with parallel filesystems, like Lustre, pNFS and FraunhoferFS (FhGFS) and is hence in particular interested in developing and providing innovative solutions build on top of pHDF5/pNexus. Additionally, we thereby hope to accelerate the maturation of open source parallel filesystems, since development of existing solutions is increasingly hampered by trends towards commercialization.

State of the Art at ELETTRA

ELETTRA (http://www.elettra.trieste.it) is a national laboratory located in the outskirts of Trieste (Italy). Its mandate is a scientific service to the Italian and international research communities, based on the development and open use of light produced by synchrotron and Free Electron Lasers (FEL) sources. The light is now mainly provided by a third generation electron storage ring, optimised in the VUV and soft-X-ray range, operating between 2.0 and 2.4 GeV, and feeding 24 light sources in the range from few eV to tens of keV (wavelengths from infrared to X-rays). The light is made available through a growing number of beamlines, which feed several measuring stations using many different and complementary measuring techniques ranging from analytical microscopy and microradiography to photolithography.

The new fourth generation light source FERMI@Elettra that is now in development is a single-pass FEL user-facility covering the wavelength range from 100 nm (12 eV) to 10 nm (124 eV). The spectral brightness available on most of ELETTRA's beamlines is up to 10^{19} photons/s/mm^2/mrad^2/0.1%bw and the peak brightness of the FEL sources is expected to go up to 10^{30} photons/s/mm^2/mrad^2/0.1%bw. The advent of femtosecond lasers has revolutionized many areas of science from solid state physics to biology. This new research frontier of ultra-fast VUV and X-ray science drives the development of a novel source for the generation of femtosecond pulses.

At ELETTRA each beamline has its own acquisition system based on different platforms (java, LabVIEW, IDL, python, etc.). This is a compromise between flexibility, feasibility and usability, allowing the scientist to autonomously maintain their application. To offer a uniform environment to the users where they can operate and store data, ELETTRA has developed the Virtual Collaboratory Room (VCR) that, among other things, allows users to remotely collaborate and operate the instrumentation. This system is a web portal where the user can find all the necessary tools and applications; i.e. the acquisition application, the data storage, the computation and analysis, the access of remote devices and almost everything necessary for the completion of the experiment. The system implements an Automatic Authentication and Authorization (AAA) based on the credential managed by the Virtual Unified Office (VUO). The VUO web application handles the complete workflow of the proposals' submission, evaluations, and scheduling. The system can provide administrational and logistical support i.e. accommodation, subsistence, access to the ELETTRA site.

The integration to the low level control system is open to various standards: BCS (the in-house control system for the ELETTRA beamlines), Tango, Grid. Thanks to the participation in many EU FP6 projects in the Grid field ELETTRA has acquired the know-how to integrate instrumentation to the Grid using the new component "Instrument Element" (IE) that was introduced by the GRIDCC project and which is now maintained and extended on the DORII FP7 project. ELETTRA hosts a Grid Virtual Organization (including all the necessary VO-wide elements like VOMS, WMS, BDII, LB, LFC, etc.) and provides resources for several VOs. The current effort is on porting many legacy applications to a Grid computing paradigm in an effort to satisfy demanding computational needs (e.g. tomography reconstruction).

State of the Art at SOLEIL

The synchrotron SOLEIL (http://www.synchrotron-soleil.fr) is a 3rd generation synchrotron radiation facility in operation since 2007. In 2009, 1,719 users have performed 348 experiments on the 14 first open beamlines. Currently, SOLEIL is delivering photons to 21 beamlines with a current of 400 mA in top-up mode: 17 beamlines are open to users and 4 under commissioning. In addition, new challenging beamlines are under construction or under design. More than 2,000 users from France, Europe and other countries are expected per year to perform experiments in various fields as surface and material science, environmental and earth science, very dilute species and biology.

On the Computing and Controls side, a great effort has been made very early to standardise hardware and software, keeping in mind developments reusability and easy maintenance. The control and data acquisition system of each beamline and the Machine control system are based on the TANGO system, initially developed by the ESRF. Since 2002, SOLEIL is very largely involved in the international TANGO collaboration which now includes five of the PaNdata partners.

Experiments carried out at SOLEIL generate datasets ranging from a few kilobytes to several gigabytes per day. All beamlines can automatically generate data in the NeXus standard format, in order to ensure easier data management and contributing to future interoperability with other research facilities. NeXus files are stored via the storage infrastructure managed with the Active Circle software, handling data availability, data replication on disks and tapes, lifecycle management. Data are accessible from the beamlines as well as from any office in the buildings, with security based on LDAP authentication. A remote access search and data retrieval system (TWIST, https://twist.synchrotron-soleil.fr) allows users to perform complex queries to find pertinent data and to download all or only parts of a NeXus file. Up to now, more than 700.000 NeXus files have been produced at the beamlines.

Data post-processing is handled either on local PCs, or on a local compute cluster dedicated to the beamline (if required for experiment control), or on the central HPC system (directly accessible from the beamlines by all the users or from any office by SOLEIL scientists). To allow data analysis applications to access experimental data independently from file format type and organisation, a generic Common Data Access API is being developed. Its is now routinely used as a unified data access layer for all our data visualisation and data analysis applications: as first result, a SAXS data reduction application is able to process SOLEIL NeXus data files as ESRF EDF data files.

For the user management and proposal submission, SOLEIL uses a revamped version of PSI's DUO, called SUNset.

SOLEIL sees this proposal as a continuation in the standardization effort, allowing for more efficiency for the scientists as well as for infrastructure managers, thanks to the development of new tools easing user management, data file exchange and access.

State of the Art at HZB

The Helmholtz Zentrum Berlin (HZB, http://www.helmholtz-berlin.de) is operating two large scale scientific facilities with an emphasis on studies on the structure and function of matter: The storage ring BESSY II is at present Germany's largest third generation synchrotron radiation source and emits extremely brilliant photon pulses ranging from the long wave terahertz region to hard x-rays. The research reactor BER II delivers beams of thermal and cold neutrons for a wide range of scientific investigations, in particular for materials sciences. The HZB also operates the Metrology Light Source, a specialised storage ring for the Physikalisch-Technische-Bundesanstalt (in Berlin-Adlershof).

On the synchrotron 46 beamlines at the undulator, wiggler, and dipole sources cover a many-faceted choice of measuring stations. The combination of brilliance and photon pulses makes BESSY II the ideal microscope for space and time, allowing resolutions down to femtoseconds and picometers. On the reactor a total of 24 measuring stations are in operation, in combination with highly specialised equipment for the most sophisticated conditions (high magnetic fields, low temperatures, high pressure). Major upgrades like NEAT-II and new stations like the High Field Magnet are currently being build.

Currently many activities focus on merging the technical and scientific support of the centre, in order to provide a more homogeneous and more effective work environment for it's users. To this end the HZB also welcomes and participates in national and European initiatives for instance within the HDRI, PaNdata, NIM3, ESRFUP and EuroFEL work packages. There is a long tradition to develop experimental stations in collaboration with external research group and other facilities, both on a scientific and technical level.

One key activity is on providing a standardised user portal for users of the BER II and BESSY II sites. as its predecessor the software is based on DUO-II standards and has been developed in collaboration with other facilities and corresponding activities within IRUVX-II. It is therefore hoped that concepts from PaNdata and HDRI activities are can be integrated as they arise.

Data management and data access procedures are not strictly standardised. It is planned to develop these further along the concepts arising from PaNdata & HDRI. The HZB for some years has had active involvement with the NeXus International Advisory Committee, and has started to establish NeXus as a data-format though more coordination in particular with other facilities will be necessary to exploit the benefits of a joint data format.

EPICS is the predominantly used control-system for the operation of the storage ring and intermixed with other technologies for the control of beamline specific devices. The HZB is an active contributor to the EPICS-community and in constant exchange with other large scale facilities using a similar approach. First concepts to extend the use of this technology to the neutron experiments are planned for the near future.

Due to the large scope of sciences covered and the strong involvement of external research groups, data acquisition systems vary throughout the site, although most experimental stations are based on in-house software (EMP/2, CARESS) and associated data acquisition hardware. Other software has been integrated into the setup, in particular SPEC and LabVIEW based systems, but also other software packages from other sites and commercial software systems.

With many of the HZB's users also visiting other facilities, the joint developments in the field of data-analysis software are welcome. Relevant work has been taken up within HDRI and will be complemented by activities like PaNdata.

State of the Art at ALBA

The ALBA synchrotron (http://www.cells.es) is currently under construction and will be fully operational in 20110. In line with this planning, the Linac and the Booster are commissioned and the storage ring commissioning will start on the 20/11/2010. The booster has reached its nominal energy on the 3/10/2010. The construction of the 7 phase-one beamlines is making good progress and the first beamline will see synchrotron light in January 2011.

The accelerator and beamline control system is done with Tango, Sardana, and Taurus based on C++ and Python for the software and on PCI, cPCI, and PLCs for the hardware. The low level control and the equipment protection system is based on PLCs from B&R. The Personal Safety system is also PLC based but on safety PLCs from Pilz. The experimental data is stored centrally on ultra-high performance disk storage (minimum 300 MB/s per client for up to 4 simultaneously writing clients on an NFS mounted disk). A server farm which will allow users to analyse their data at ALBA from their home institutes has been projected. One fast two dimensional detectors which will produce up to 80 MB/s has been purchased and the tender process for others will started soon. The user office has been created and the selection of the software approach in this area is currently carried out. The ALBA computing division is dedicated to use best practices (ITIL, Prince2, ...) and centralized automated tools (e.g. the network, TANGO, PLC programs and GUIs are automatically created out of the cable data base).

ALBA is actively participating in the TANGO collaboration and is leading the development in the new generic data acquisition system Sardana in collaboration with the ESRF and DESY and possibly MaxLab.

Being in the commissioning phase, ALBA will not be able to participate in the software developments proposed within the PaNdata project to the same extend as some of the more mature institutes. ALBA will follow the ongoing discussions, participate in the policy, and dissemination and development activities, and will readily deploy the outcome of the PaNdata developments.

State of the Art at CEA/LLB

The Laboratoire Léon Brillouin (**http://www-llb.cea.fr**) is the French national neutron scattering facility located at Saclay in the nearby of Paris. It operates 25 instruments distributed around the Orphée reactor which is operated by the CEA.

The instruments are very heterogeneous and produce data amounts ranging from few kB to few hundreds of MB a day. They are individually controlled by various type of software using different platforms, either developed by the IT department or the instrument responsible.

The data acquisition is specific for each instrument and there is no standard for data format at the LLB, but the majority of the data storage is performed either on XML or more recently NeXus type format. The XML format is preferred by the user community because of the simplest possibility for visualising, and some of the instruments store data on both XML and NeXus formats. The raw data are stored on each instrument and the storage policy of the LLB includes copy of the data onto two different central repository located in physically different places. The data are stored without limit in time and are internally freely accessible. All the data are available to anybody on request.

Data post-processing is handle by each scientist on his own computer. Some software suite have been developed by the scientists of the LLB which are available freely on the website of the laboratory, other scientists use software available from other institutions.

The Laboratoire Léon Brillouin observes very favourably the development of common data formats for the different IR because it will help to improve the synergy between the different facilities, by promoting the scientific study combining several experimental techniques. We hope that we will also benefit from common software developments, and an easier access of the different facilities.

1.2.2 State of the art of the Technology

This section reviews some current initiatives in data policy and then goes on to review the state of the art of the technology from the perspective of the three Joint Research Activities of this project: supporting provenance, supporting long-term data preservation, and scalable data flow frameworks.

1.2.2.1 Managing data within Facilities

PaNdata Europe Strategic Working Group (PaNdata Europe) Support Action is currently establishing common a policy framework across the partners. This policy framework provides an excellent basis for developing an Open Data Infrastructure (ODI). The first step towards the ODI will be the implementation of these policies at the partners' facilities. Implementing these data policies will have a number of beneficial consequences in that experimental data will become openly available with scientists fully aware of the rules and limitations.

Prior to the concerted activities of the PaNdata consortium, long term archival of photon science data was a rarity. The mere existence of a draft data policy had already a positive impact. PNI-HDRI for example will implement a data policy which conforms with the recommendation of the Helmholtz society. Though it hasn't been ultimately decided yet, there is little doubt that the policy developed by PaNdata will be implemented. However, implementing the policy framework and further advertising it beyond PaNdata remains an important activity.

Archival of and access to scientific data can be greatly facilitated by standardizing associated metadata and data formats. As proposed in the data policy, and already been implemented at several neutron and photon sources, Nexus/HDF5 will provide a suitable standard, also recognizing the recommendation of the European Commission to use HDF5 for all binary data. To promote the standardization communication between facilities, application developers, detector vendors and user communities will become increasingly important. One suitable action is the stronger participation in standardization bodies like the Nexus International Advisory Committee (NIAC). PaNdata ODI can create the momentum required to direct application developments and implementations towards the chosen standard. Such a momentum can be further enhanced liaising with other facilities, like the European Spallation Source (ESS) or the European XFEL, strong contacts already exist, for example through the participation of PaNdata partners in ESFRI projects like CRISP.

CRISP, the proposed ESFRI Cluster project for the physical and material sciences, concentrates on the technological basis to securely transfer data from the instrument towards a storage element. EuroFEL develops an authentication and authorization infrastructure (AAI). Other projects tackle the data continuum more from the side of specific experiments and applications. PaNdata, who is participating directly or indirectly in many of these projects, can bridge the gaps between the different approaches by intensifying the cooperation. PaNdata partners have already ensured that different aspects of CRISP and PaNdata ODI remain fully compatible and complementary, for example with respect to investigations of filesystems, standards or AAI.

Photon and Neutron science facilities are occasionally being transformed or upgraded to a "next-generation" instrument, but hardly ever vanish from the scientific landscape. It is hence less difficult for a single facility to ensure sustainable data infrastructures and to encourage scientists to participate in data curation. However, such an infrastructure will be considerably more valuable if it is federated between facilities, which requires close cooperation between

the facilities as well as common data sharing policies and standards. The policy framework developed by PaNdata Europe provides a sufficient basis to achieve a smooth integration of data repositories across facilities. Furthermore, by selecting a widely accepted standard data format and data catalogue, cross-disciplinary utilization of scientific data can be greatly facilitated, overcoming the obstacles of highly domain specific data repositories. Promoting the Open Data Infrastructure to other related projects might well broaden the interoperability and inter-disciplinarily of scientific data repositories.

1.2.2.2 Provenance Support

The path from a scientific proposal, through performing an experiment at one or more facilities, then numerous stages of data analysis and derivation, and ultimately to the publication and post-publication validation of the results, is a highly iterative and interactive multi-stage process. This process is typically a collaboration between academic researchers, possibly from several institutions, and facility based staff and takes place in part at the researchers home institution and in part at centralised facilities.

In recent years, the PaNdata consortium and others have worked progressively towards supporting the data continuum from application to publication. For reasons of expedience, this has been particularly focused on those parts of the process where the facilities have a high level of control. In particular, the pathway from the two ends of the data continuum towards the centre. On the one hand, working forwards, systems have been developed which carry information from the proposal towards the experimental equipment and use this to tag data collected together with metadata, which is used in cataloguing and archival. On the other hand, working backwards from publications, trawling and annotation with metadata about the facility, instrument and experiment, and linking to the data collected at the start.

The proposed project will concentrate on the core of the data continuum, the data processing framework, where research teams perform the analysis of the experimental data tightly interlinked with an immense library of external information systems. This central part of the process, which currently is not integrated into the data infrastructure to the same degree, is more challenging for several reasons. Firstly it is in less prescriptive in nature, often proceeding by a process of trial and error, tuning parameters and employing a variety of software tools, and combining different data sources. Secondly it is often performed off site and in most cases at several cooperating sites. However, capturing and recording the data processing pipeline is absolutely essential if the provenance of the analysed results is to be established. Large facilities, such as those in PaNdata, are ideally placed to provide a lead in developing such support for the science which uses those facilities as they are highly trusted by their communities, have dedicated staff in place, and can provide reliable central infrastructure which is independent of any particular research group.

The PaNdata partners, through the ongoing Support Action, have already met (or are about to meet) the preconditions for building a rich support environment for its user community across the whole of the data continuum. We will have defined:

- a common data policy framework;
- a protocol for exchange of user information;
- a definition of standards for common scientific data formats;
- a strategy for the interoperation of data analysis software; and
- a framework for the integration and cross-linking of research outputs.

Furthermore, the ICAT[2] provides a catalogue of experiment information and raw data as the starting point of capturing the relationships between derived scientific objects.

Thus the PaNdata consortium is now in the position to be able to extend the existing infrastructure to cover the data processing framework and so support its scientific community across the whole data continuum, to manage the provenance of data. This infrastructural support for the data analysis process will require a number of components which are described below.

1.2.2.3 Modelling the Data Continuum

A core component of the infrastructure is an information model for the whole continuum, so that the stages of analysis can be recorded and traced. The Core Scientific Metadata Model, CSMD[3], is becoming accepted by the Photon and Neutron community as a core model to catalogue data, and forms the central model of the ICAT software suite. To integrate the data analysis within the data infrastructure framework, this model needs to be extended to cover the data reduction pipeline. Within the UK, the Integrated Infrastructure for Structural Science (I2S2) project[4] is developing such a model which extends the CSMD with support for recording stages of scientific analysis defined in the ORE-CHEM project[5]. A key feature of this approach is that provenance relationships can be captured across institutions and databases. This work is compatible with the Open Provenance Model which draws together the research into provenance which has been undertaken in the last decade.

In this project, we will bring this exploratory work into practise via further enrichment of the abstract provenance model into a practical environment; integration into the analysis tool environment, so that the provenance can be captured during the execution of analysis; evaluation in different scenarios across the consortium, and improved tool support. We will collaborate with the successors to the W3C Incubator Group on Provenance to ensure that our approach was compatible with the emerging standards in the area.

Note that the approach taken here is not the one used in many scientific workflows, such as Taverna, Kepler, YAWL, and enacted via languages such as BPEL. The data analysis activities are not predefined to a script, but often follow an exploratory approach, with freedom for the scientist to take different routes. Thus the model, and the framework which implements it, has to be concerned with registering, linking and logging the work done, rather than steering it. This then allows for the tracing of connections between data, including the analysis codes used, so that the significance and provenance of any artefact can be retrieved for the scientist themselves and the wider community.

1.2.2.4 Modelling the experimental context

The above generic model needs to be instantiated to various experimental techniques. Consequent utilization of a well structured and object-oriented implementation is the most efficient way to construct an experiment specific layout from a generic instrument description, which is a particular strength of Nexus' HDF5 implementation. Within PNI-HDRI, some PaNdata partners have thus begun to create an abstract definition of a beamline, as a blueprint to fully re-construct an experiment at a particular beamline from generic building blocks (aka application definitions in Nexus terms) which will serve as an initial

[2] http://code.google.com/p/icatproject/

[3] http://epubs.stfc.ac.uk/work-details?w=51838

[4] http://www.ukoln.ac.uk/projects/I2S2/

[5] http://research.microsoft.com/en-us/projects/orechem/

scheme for facility and instrument ontologies, complementing exploratory work at STFC on developing an ontology for ISIS.

IN this project we will develop an ontology framework for facilities, with general classes spanning across facilities; specific specialisations of the framework for individual facilities; and integrate functionality for indexing and searching with ontologies within the tool support, such as the ICAT data management suite and its front end TopCat. Thus this will represent a significant extension of the current best practice within the communities the facilities support.

1.2.2.5 Sharing

Data sharing has been recognised as a key driver for scientific advance in the twenty-first century, notably in the "OECD Principles and Guidelines for Access to Research Data from Public Funding". This is being continued in the PaNdata Support Action work on data policy. This needs to be developed further by integration into the tool support. The ICAT front-end system TopCat allows common search and access to data holdings in different facilities; this will be developed further and rolled out to more participating facilities within PaNdata.

1.2.2.6 Supporting Preservation

Awareness of threats to long-term preservation of digital assets of all kinds is growing. The PARSE.Insight project found that data managers in many different communities are highly aware of the threats to long-term preservation of their data. In the survey conducted by the project, the top three threats regarded as either important or very important were: lack of sustainable hardware, software or support (86%); problems with understanding the semantics, formats or algorithms of data (83%); uncertain origin and authenticity (81%).

The same project produced a roadmap for the development of a science data infrastructure focussed on long-term preservation. The roadmap comprises aspects that are financial, organisational/social, policy-related and technical.

1.2.2.7 Preservation Frameworks

PARSE.Insight worked within the framework of OAIS (Open Archival Information Systems), the international standard[6] for organisations with the responsibility to preserve information and make it available. OAIS proposes a reference model to capture the capture the stages of the process of preserving digital information in the form of Archival Information Packages (AIP). Its data model discusses the components of the AIP, augmenting the data object itself with a variety of Representation Information (RI) that is information items which describe the context in which the data needs to be interpreted. Such RI could be a description of syntax (a data format), of semantics (an ontology, or user guide), or of processing (interpretive software). A key notion is that of Designated Community, which capture the assumptions we can make of the knowledge of the target community of the preservation activity.

A number of recent projects have brought this conception closer to practice. The Planets project concentrated on preservation planning within national libraries and national archives; this work is being extended to explore scalability issues, including those for scientific data in the SCAPE project. The Planets project has also lead to commercial tool support, such as Tessala's Safety Deposit Box. The CASPAR project developed tool support for the OAIS framework including RI, packaging, designated community, and authenticity.

[6] public.ccsds.org/publications/archive/650x0b1.pdf

Further initiatives such as the Alliance for Permanent Access to the Records of Science and the UKs Digital Curation Centre (DCC) have raised the general level of awareness of preservation issues. A particularly significant issues is the cost/benefit of digital preservation; the cost of preservation may not outweigh the long term benefit, and a systematic approach needs to be taken to evaluate this. This aspect has been explored in CASPAR and within the UK DCC.

Science data has certain particular requirements for long-term preservation, especially connected with semantics. It is important to be able to correctly interpret the data in future - for example, the units of measurement, the precision, implicit knowledge about the data taking process. Further, there is a need to maintain the context of the whole data environment, including the analysis software used; this software would also need preservation actions, a complex issue only recently explored in for example a number of recent UK JISC projects[7].

1.2.2.8 Preservation within Facilities

PaNdata facilities have in the past not seen their primary function as the holders of data for the long term; data has been seen as relatively ephemeral, taken away by the user and not necessarily maintained at the facility. However, the much larger data volumes has meant that large central data storage is needed to hold the data (it cannot be taken to the users institution on a disk). Further, there has been recognition that data needs to available for reuse in the future to check the validity of science research and also maximise the value to be gained from reanalysis of expensively acquired data. Consequently, facilities have developed systematic approaches to data management and storage, and are exploring the cost/benefit of preserving data for their communities. Thus developing a framework for preservation, specialised to the PaNdata facilities and their user communities, represents a significant advance on the current state of the art.

There are several strands to enabling preservation and reuse of data holdings.

Cost/Benefit Analysis. Based on initial work within the current support action, a systematic analysis of the costs and benefits of digital preservation will be undertaken, applying the work of CASPAR and the DCC to identify those data which would gain the most benefit from preservation and develop framework for preservation.

Create and maintain representation information. With a diverse range of disciplines, methods and tools supported by facilities, if data is maintain its reusability over time, its context needs to be maintained, which can be captured by Representation Information (see below). Currently, facilities do not generally maintain a systematic approach to capturing such context.

Supplementary information is needed to allow the interpretation of datasets (for example, units of measurement, knowledge about the conditions under which the data was captured, uncertainties or errors), and thus Representation Information is required. In the context of scientific data, one kind of supplementary information is offered by the papers written on the basis of the data (whether final peer-reviewed publications or "grey literature"). Indeed it has been said that the best metadata about a dataset is the papers written about it. This implies that the whole research lifecycle can be employed as a source of representation information. In this project we will provide systematic support for digital preservation across the research lifecycle

Authenticity of a digital object. The integrity of facilities data needs maintained over time, encompassing not only checking bit-level integrity but also the mechanisms for ensuring that

[7] http://www.e-science.stfc.ac.uk/projects/software-preservation/preserving-software.html

threats to integrity are controlled, for example by enforcing access policies. While facilities generally have strong support in place to manage and store data safely over time, this needs to be certified so that the authenticity of the data can be assured.

Digital rights. Facilities support a diversity of users, funded from different sources, so that there is a complex ownership situation. These digital rights need to be clarified in policies, this is currently being undertaken within the Support Action. These need to be enforced within a preservation framework to ensure that the appropriate rights are respected in the future.

Persistent identifiers. In order to maintain a stability of reference for data sets generated within facilities, a persistent identifier needs to be used to identify them. Currently facilities are exploring the use of persistent identifier schemes, for example the ISIS facility is adopting the Digital Object Identifier (DOI) scheme to ensure that the citation of objects is maintained. This will be extended into derived data which may require use of persistent identifiers at finer granularity.

Software Preservation. The current support action is working towards a registry of data analysis software. This would be enhanced by the addition of metadata to support preservation, and making it available for use as representation information in the PaNdata preservation framework.

1.2.2.9 Supporting Scalable Data Flow

Complex experiments at Synchrotron or Free Elecron Laser facilities frequently require accumulation of data from several detectors in parallel, which becomes increasingly challenging with framerates in the kHz to MHz range, resulting in data generation rates which of up to 40-100GB per hour. . Real-time analysis and data reduction requires parallel processing of the events, which could possibly be implemented on specialized hardware like for example on a platform combining CPU, GPUs and/or FPGAs.

All current implementations rely on strictly sequential read/write access to the raw experimental data, even in the case of multi-threaded IO. Taking up and accumulating parallel data streams into a manageable number of digital objects is hence an open issue.

The HDF Group, responsible for the definition of the Hierarchical Data Format (HDF), has recently released a parallelized version of HDF5 (pHDF5). It builds up on MPI-IO and parallel filesystems. HDF5 has been proposed as an standard for binary digital objects by the EC, and the PaNdata consortium has selected HDF5/NeXus as the recommended standard for Photon and Neutron Science experimental data. Implementation of HDF5/Nexus for specific applications is an ongoing activity which is progressing well. HDF5/Nexus is hence the only suitable standard to build on the development of a parallelised solution to the data challenge.

Implementation of pHDF5 in the Nexus API would be highly beneficial for both the facilities as well as the scientific user communities. This requires careful examination of the underlying infrastructure like parallel filesystems and protocols (e.g. Lustre, pNFS4.1, FhGFS, PVFS to name a few), optimization of parallel data stream generating engines like GPUs or FPGAs and demonstration of the capabilities on specific use cases.

This approach is complementary to the work proposed by the Physical Science ESFRI Cluster project CRISP, which intends to investigate filesystems and protocols with respect to extreme datarates, and investigation of GPU-based (pre-)processing of raw experimental data as currently being investigated by the PNI-HDRI project of the Helmholtz society (HGF). The proposed work would nicely close the gap between the more hardware, caching and protocol oriented investigation of CRISP and PNI-HDRI on one side, and the development of applications based on standardized HDF5/NeXus experimental data on the other side.

The proposed virtual laboratories (WP5), both crystallography and tomography, will also serve as demonstration show cases for the implementation of pHDF5/pNeXus under real experimental conditions. Data rates and data volumes generated at advanced facilities like the European XFEL will exceed those at conventional synchrotron light sources by at least an order of magnitude; estimates are of up to 7TB per hour, depending on the experiment. Investigation of the requirements and hardware implementations necessary to cope with these extreme conditions is an entirely different topic. However, the developments within this workpackage can serve as a proof of principle to guide developments for x-ray free electron lasers coming up in future years.

Although the stability and performance for high data rates and high performance analysis will depend to some extent on the filesystems, the implementation in NeXus will be fully portable and protocol independent. The approach can therefore easily extended towards distributed filesystems like WebDFS, which would allow a research team to coherently analyse data collected and residing at different facilities within a single pNeXus-supporting application, though the available network infrastructure might pose certain limitations for certain types of applications.

In such cases, sophisticated data transfer mechanism will become unavoidable. However, at present there is a functional chasm between the facility and the visiting scientist's home institute. Data needs to be transferred using traditional physical or networked media which breaks the relationships between the items, thus causing a loss of context. In order to increase their effectiveness, researchers need to move data across institutional and domain boundaries in a seamless and integrated manner: this workpackage seeks to “bridge the chasm” and develop a robust framework to enable these seamless transformations to take place routinely and which will greatly increase researcher efficiency and productivity. There will also be greater return on investment in the central facilities through more cost-effective use of resources by the client base. We will enable seamless management, via an overarching framework, that will allow researchers to simply and efficiently manage their data across institutional and administrative boundaries.

In both cases, the proposed development of a common authentication and authorization framework and a strong collaboration between the facilities within PaNdata and beyond will be highly beneficial.

1.3 Methodology to achieve the objectives of the project, in particular the provision of integrated services

1.3.1 Structure

The workplan is directed towards the development, deployment and operation of a suite of common technology for the management of data at the participating facilities. This technology will support a set of integrated services that provide transparent access for users across participating facilities to a common catalogue of curated and provenanced data with progress towards developing the capability to deliver that data at rates expected from the next generation of instruments.

The deployment and operation of these common services across the participating facilities requires *coordination* of activities within the project and externally, some *research and development* to enhance generic technologies to the specific environment and integrate with existing deployed systems, as well as the *deployment and operation* of actual services to specific communities.

The project is broken down into 8 workpackages which together cover the objectives given in Section 1.1.5 above. Workpackages 1 and 2 are Networking Activities specifically dealing with management of the project and engagement with other related initiatives and cover objectives 1 (collaboration) of section 1.1 above. Workpackages 3 and 4 are Services Activities and cover the deployment and operation of the user and data catalogues described in objectives 2 and 3 (Users, Data). Workpackages 6 and 7 (Provenance and Preservation) are Joint Research Activities covering the research and development of technology required to ensure the services are supported by technology which can deliver high quality data. Workpackage 8 (Scalability) will undertake research and development of technology to ensure that the services can use the data infrastructure effectively to deliver data at the rate necessary to support the next generation of detectors and data acquisition systems. Finally, Workpackage 5 (Virtual Laboratories) is a Service Activity which will work with users of three specific techniques to adopt and evaluate the services. Together these workpackages will deliver the functionality required for end to end support of the data continuum "*from application to publication*".

Networking Activities	
1	**Management** of the project and related internal activities.
2	**Engagement** with related external activities.
Service Activities	
3	Deployment, operation and evaluation of a common **data catalogue**
4	Deployment, operation and evaluation of a common **AAA/users catalogue**
5	Deployment of **Virtual Laboratories** serving particular techniques
Joint Research Activities	
6	Research & development of shared technology to track the **provenance of data**
7	Research & development of shared technology to ensure the **preservation** of data
8	Research & development of shared technology to enable the **scalability of** data transfer

The four core technological workpackages, 3, 4, 6, and 7 (users, data, provenance and preservation) continue the standardisation work being undertaken in the four themes of the current PaNdata Support Action (see figure below).

For all 4 streams, standardisation at policy level is already being undertaken through the PaNdata Support Action and this project will build on that work to implement its services. For Users and Data, the relevant technology is already mature enough for deployment and operation to be now implemented through a service activity (Workpackages 3 and 4). On the other hand, for streams on provenance and preservation, there is still research and development required to develop technological components which can be incorporate into the services. This will be undertaken through Workpackages 6 and 7. Furthermore, it is clear that enormous data rates predicted for the next generation of instruments makes it timely now to develop protocols which can effectively use the future data transfer infrastructure as developed in other projects (e.g. GEANT2, EGI, EGEE-III) in an effective manner.

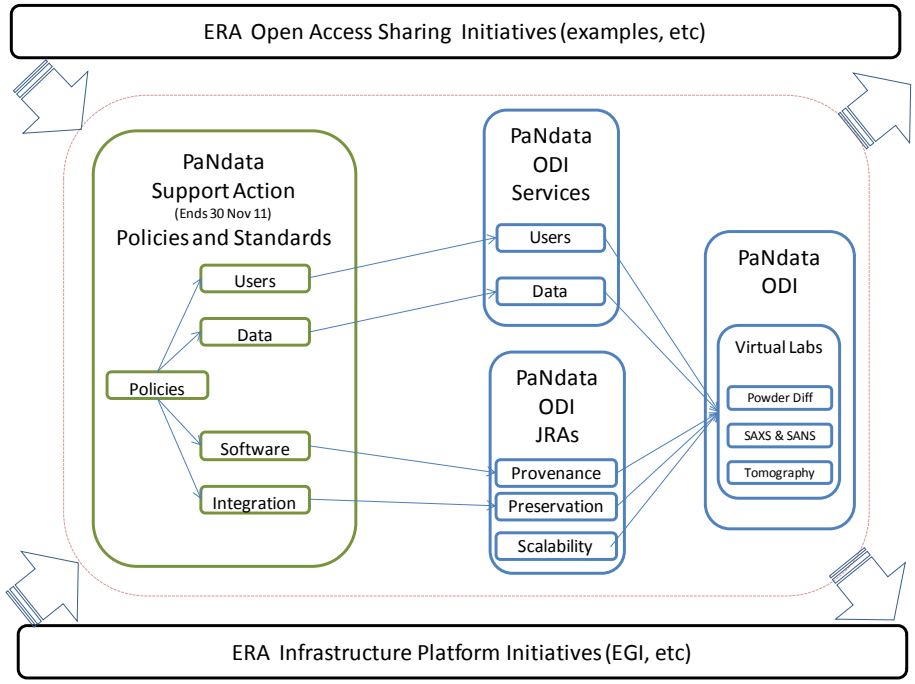


Fig. 5: PaNdata ODI builds on four areas of standardisation from the PaNdata Support Action

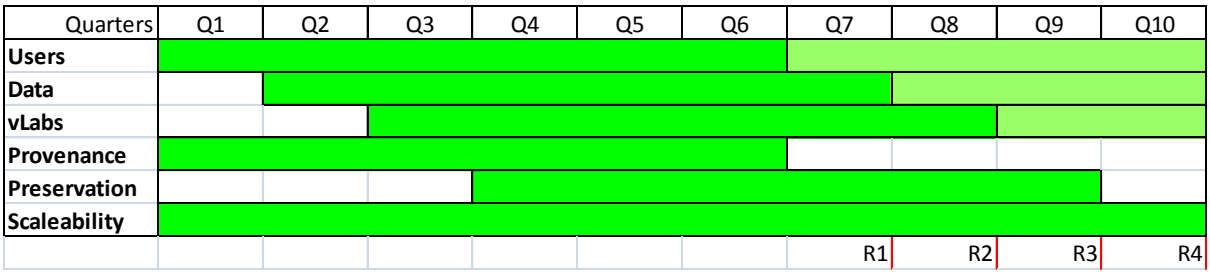
1.3.2 Schedule

Although the four streams have dependencies between them which constrain their scheduling, for example the ability to share data from the catalogue clearly requires common authentication across facilities, the workpackages are reasonably independent so some load balancing is possible whilst remaining consistent with the overarching aim of establishing the user and data services sufficiently early to enable operation and evaluation in the virtual laboratories within the time span of the project.

The overall duration of the project is set at 30 Months which although ambitious is achievable because the consortium is already working together effectively and the policy and other prerequisite work is already underway. However, a staggered start to the workpackages will enable attention to be focused on each topic to ensure the work gets underway progresses quickly.

Whilst JRAs will develop technologies that enhance the functionality, the deployment of the service deployment will not be held back until this technology is ready. Rather a quarterly

release cycle will be implemented during the last year of the project where enhancements are added to the basic functionality as they become available.



The major development period for each workpackage and service release points

After the completion of the development of the new services, their operation will become internalised into the ordinary operating procedures of the facilities and an evaluation undertaken. This is indicated by the lighter shading in the diagram above.

1.3.3 Milestones

Milestones are used in this proposal to mark the major stages of the project development, rather than individual handovers between workpackages. The major project milestones are at months: 9, 21 and 30. These stages mark:

M1. The completion of the detailed definition of the user AAA and data catalogue services which drive the research in the provenance and preservation, and the definition of the virtual laboratory usage scenarios which set requirements for the scalability research.

M2. The completion of the first release of user AAA and data catalogue services and second delivery of research outputs in all three JRAs. This forms the baseline for integration of all workpackages which takes place in the last 9 months of the project.

M3. The completion of the project with reports of the evaluation of the integrated services.

The work packages and milestones are described in more detail in sections 1.4, 1.5 and 1.6.

1.3.4 Dependencies

Key dependencies in the project are as follows:

- The completion of policy and standardisation work being undertaken by the consortium in the PaNdata Support Action. The support action ends in November 2011 however the key deliverables are all scheduled at least 2 months before that.
- The establishment of the shared user AAA service will be required to underpin the integrated data catalogue and both of these will be required to enable seamless access to the content through the virtual laboratories.
- The traceability of analysis delivered through the provenance JRA will be incorporated into the first release of the services whereas the preservation framework will not be available until the third release.
- The scalability work has no prerequisite and will be integrated into the services at the end of the project.

The dependencies within work packages are described in more detail in sections 1.4, 1.5 and 1.6.

1.4 Networking Activities and associated work plan

The Networking, Service and Research activities in this I3 project are best understood in the context of the project as a whole. For this reason, several tables in this section describe the work plan for the whole project and are repeated verbatim in the sections 1.5 and 1.6 with grey shaded sections to highlight the relevant part. The table below summarises the scope of each subsection.

Section No.	Describes	Scope
1.4.1	Overall strategy of work plan	**Network Activities only**
1.4.2	Timing of the different WPs (GANTT)	Whole project
1.4.3	Work package list	Whole project
1.4.3	Deliverables list	Whole project
1.4.3	Description of each work package	**Network Activities only**
1.4.3	Summary effort table	Whole project
1.4.3	List of milestones	Whole project
1.4.4	Graphical presentation of components and interdependencies (Pert)	Whole project
1.4.5	Risk analysis for service activities	**Network Activities only**

Scope of description of each subsection within this section

1.4.1 Overall Strategy for Networking Activities

The overall strategy of the work plan for the whole project is described in Section 1.3. This section describes only those aspects which are specific to the Networking Activities.

The Networking Activities address those elements of the project which relate to managing the collaboration and coordination of activities. The Management workpackage relates to the coordination of activities within the project and the Engagement workpackage relates to coordination of activities with activities and initiatives outside the project. The two Network Activities cut across the technical activities of the other workpackages.

1.4.2 Schedule

Quarters	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Milestones			M1				M2			M3
Network Activities										
Management	D			D				D		D
Engagement	D				D		D		D	
Service Activies										
Users		D		D		D		D		
Data			D		D		D		D	
vLabs		D			D			D		D
Joint Research activities										
Provenance				D		D		D		
Preservation					D		D		D	
Scaleability			D				D			D

Schedule of workpackages for PaNdata ODI

Key.

D mark the quarters in which workpackages have major deliverables.

M1-M3 are the project milestones.

For clarity, dependencies are not marked here but described in the Pert chart later.

The lighter shaded area in the technical workpackages corresponds to periods of time when services are integrated into the normal operations of the facilities.

1.4.3 Detailed Work Description

Workpackage list (with the grey shaded work packages of the networking activities)

Workpackage No.	Work package title	Type of activity	Lead Partner No.	Lead (short name)	Person Months	Start Month	End Month
	Networking Activities						
1	Management	COORD	1	STFC	10	1	30
2	Dissemination	NA	6	DESY	26	1	30
	Total (Networking Activities)						
	Service Activities						
3	User AAA Service	SVC	2	PSI	63	1	30
4	Data Catalogue Service	SVC	7	ELETTRA	57	4	30
5	Virtual Laboratories	SVC	6	DESY	51	1	30
	Total (Service Activities)						
	Joint Research Activities						
6	Provenance	JRA	1	STFC	36	7	30
7	Preservation	JRA	3	ILL	36	10	30
8	Scalability	JRA	4	DIAMOND	36	1	30
	Total (Network Activities)				**36**		
	TOTAL (All Activities)				**315**		

Deliverables List (with the grey shaded deliverables of the networking activities)

Del No.	Deliverable Name	WP No.	Nature (R/P/D/O)	Diss. Level	Del. Date
2.1	Project Website	2	O	PU	1
1.1	Project manag't structures, reporting, risk and quality procedures	1	R	CO	3
2.2	Dissemination plan	2	R	PU	3
3.1	Specification of AAA infrastructure	3	R	PU	6
5.1	Specific requirements for the virtual laboratories	5	R	PU	6
4.1	Requirements analysis for common data catalogue	4	R	PU	9
8.1	Definition of pHDF5 capable Nexus implementation	8	P	PU	9
8.2	Evaluation of Parallel filesystems and MPI I/O implementations	8	R	PU	9
1.2	First annual management report	1	R	CO	12
3.2	Pilot deployment of initial AAA service infrastructure	3	P	PU	12
6.1	Model of the data continuum in Photon and Neutron Facilities	6	R	PU	12
2.3	First Open Workshop	2	O	PU	15
4.2	Populated metadata catalogue with data from the virtual laboratories	4	R	PU	15
7.1	Implementation of persistent identifiers for PaNdata datasets	7	D	PU	15
3.3	Production deployment of AAA service infrastructure	3	D	PU	18
5.2	Deployment of Specification of the three virtual laboratories	5	R	PU	18
6.2	Common ontology def'n and def'n of tools to support provenance	6	R	PU	18
2.4	Open Source software distribution procedure	2	R	PU	21
4.3	Deployment of cross-facility metadata searching	4	D	PU	21
7.2	Mechanisms and tools for representation information and archiving	7	R	PU	21
8.3	Implementation of pNexus and MPI I/O on parallel filesystems	8	P	PU	21
8.4	Examination of Distributed parallel filesystem	8	R	PU	21
8.5	Demonstrate capabilities on selected applications	8	D	PU	21
1.3	Second annual management report	1	R	CO	24
3.4	Evaluation of initial AAA service infrastructure	3	R	PU	24
6.3	Tools for building research objects in Photon and Neutron Facilities	6	P	PU	24
2.5	Second Open Workshop	2	O	PU	27
4.4	Benchmark of performance of the metadata catalogue	4	R	PU	27
7.3	Mechanisms and tools for integrity of datasets	7	R	PU	27
8.6	Evaluation of coupling of prototype to multi-core architectures	8	R	PU	27
1.4	Final management report	1	R	CO	30
5.3	Report on the implementation of the three virtual laboratories	5	R	PU	30
6.4	Evaluation report on provenance management	6	R	PU	30
7.4	Report on evaluation of preservation mechanisms	7	R	PU	30

Description of each work package:

Work package no.	1			Start date or starting event:						M1
Workpackage title	Management									
Activity Type	MGT									
Part. number	1	2	3	4	5	6	7	8	9	10	11
Part. Short Name	STFC (Lead)	ESRF	ILL	Diamond	PSI	DESY	ELETTRA	Soleil	ALBA	HZB	CEA
Person-months	10										

Objectives

- To establish an effective and efficient collaboration between the partners delivering added value to each participant through shared networking, service, and research activities.
- To ensure that the project achieves its objectives with the agreed budget and time scales and to the required quality.
- To report to the Commission as required.

Description of work

Methodology:

The methodology is described in section 2.1 of this proposal, 'Management structure and procedures'. The key aspects of the methodology are:

- An appropriate structure of boards, individuals and groups with clearly defined decision making powers and responsibilities.
- Meetings and other communication at suitable frequency and with clear purpose.
- Procedures for management of quality and risks.
- Defined reporting timetable to the EC.
- The Consortium Agreement for managing relations between project partners.

Tasks

Task 1.1: Set up mechanisms to run the project through the rest of its duration (M1–M2).

Task 1.2: Monitor progress of project activities and put in place appropriate corrective actions if this progress falls short of that required to deliver the project (M1–M30).

Task 1.3: Organise general meetings of the project (kick-off and bi-annually thereafter).

Task 1.4: Report to EC on the technical and financial progress of the project (annually and at the end of the project).

Deliverables and month of delivery

D1.1 : Project management structures, reporting, risk and quality management procedures (M3)

D1.2 : First annual management report (M12)

D1.3 : Second annual management report (M24)

D1.4 : Final management report (M30)

Work package no.	2			Start date or starting event:					M1	
Workpackage title	Engagement and Dissemination									
Activity Type	COORD									
Part. number	1	2	3	4	5	6	7	8	9	10	11
Part. Short Name	STFC (Lead)	ESRF	ILL	Diamond	PSI	DESY (Lead)	ELETTRA	Soleil	ALBA	HZB	CEA
Person-months	2	2	2	2	2	6	2	2	2	2	2

Objectives

- Engagement with other initiatives and dissemination of project results, in particular to other research infrastructures.

Description of work

Methodology:

In this workpackage the PaNdata ODI will concentrate its networking activities on its key stakeholder groups, especially facility user communities, partner research institutes/organisations in Europe and world-wide, and more general (e-)infrastructure developments, within ESFRI and other national and international programmes. In particular, it will strengthen and possibly formalize cooperation with ESFRI projects EuroFEL, CRISP, ESS and European XFEL.

The project will early on build on its current communities to form an interest group of users and also infrastructure support and development personnel, via its website and other social network building activities, allowing the project to inform the community of its progress and encourage feedback and participation in the ongoing development of PaNdata. This community will be invited to participate in project workshops which will showcase the work of PaNdata and request contributions from the community to further its progress.

PaNdata currently represents a part of the Photon and Neutron communities in Europe and elsewhere, and a key objective is to extend the collaboration. Thus a key part of the dissemination plan of this workpackage will be to develop a roadmap to widen the scope of PaNdata for participation of other facilities, promoting the work PaNdata to those facilities and consulting with them on their requirements and their current best practise.

Further the project will align itself with other related e-infrastructure and data integration developments outside the project, in particular within the European Data infrastructure programme for e-Science and elsewhere across Europe, with a view to the longer term integration of this work into the broader infrastructure required to support European Research in the coming decade. Members of the consortium will actively participate in programme co-ordination meetings, contributing insights from the project and adopting best practise from other projects as appropriate. Thus the project will contribute to the development of the broader infrastructure through participation in relevant integration, planning and standardization activities required to achieve the eIRG vision of an integrated European e-Infrastructure.

Standardisation activities are also an important route for dissemination and PaNdata-ODI will participate in standardization bodies like Nexus International Advisory Committee (NIAC), W3C, ISO for OAIS, and other appropriate standards bodies to ensure the wide application and interoperability of its tools to the wider community. Further, participation in CODATA (International Council for Science : Committee on Data for Science and Technology) working groups and events will raise the visibility of the work of PaNdata to the wider scientific community.

The consortium will disseminate its developments and standardization activities to user and application developer communities, especially to facility user meetings to encourage adoption by users, and also the bi-annual NoBugs event, which targets facilities infrastructure developers, but also to wider community events, such as scientific communities, such as the International Union of Crystallography, and also technology forums such as the International e-Science conference, the Digital Curation conference and the Open Repositories conference.

Tasks:
Task 2.1. Establish an external web site as an extension to the existing website for the PaNdata collaboration (www.pandata.eu).
Task 2.2. Establish an interest group for project news items via community channels, informing them of project progress.
Task 2.3. Presentations to relevant international audiences at conferences, symposia, other project meetings etc.
Task 2.4. Provision of the open source software and appropriate documentation to potential partner bodies.
Task 2.5. Workshops to present the integrated systems to user and facility communities.

Deliverables and month of delivery

D2.1 : Project Website (M1)
D2.2 : Dissemination plan (M3)
D2.3 : First Open Workshop (M15)
D2.4 : Open Source software distribution procedure (M21)
D2.5 : Second Open Workshop (M27)

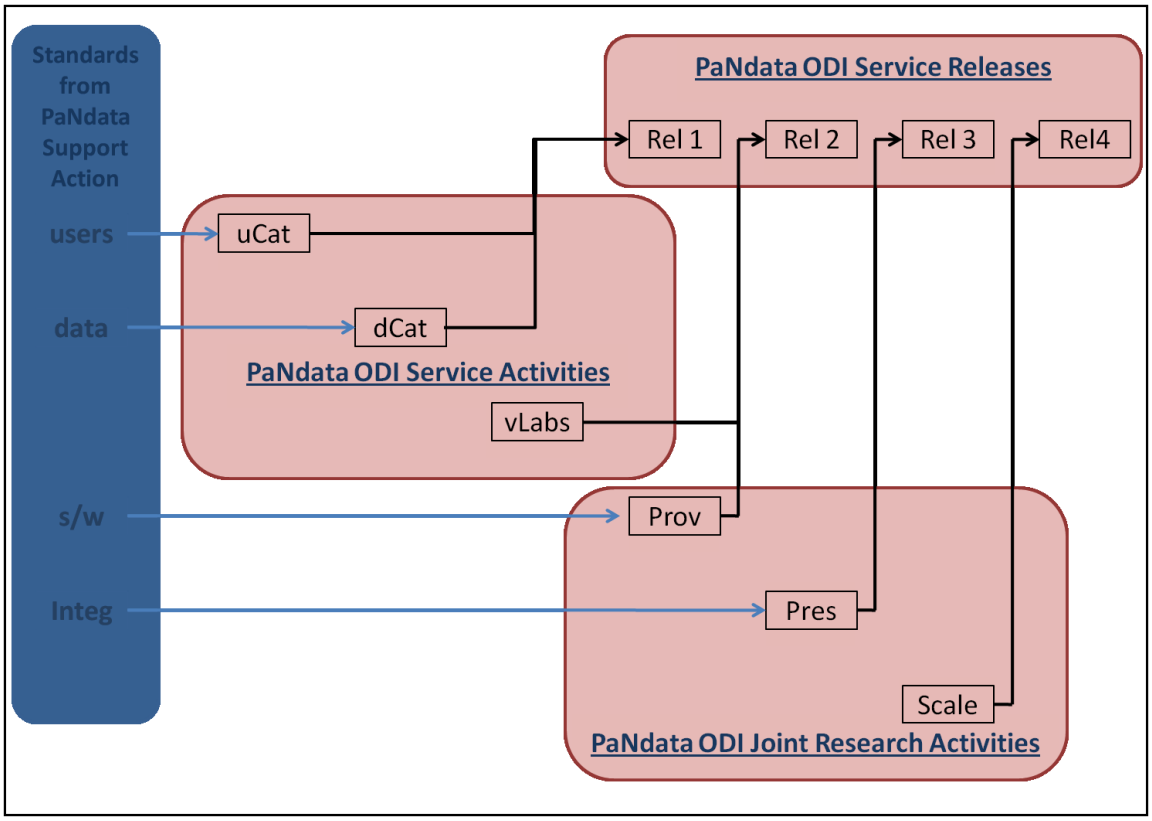
Summary effort table

Partner Number	Short Name	COORD 1	2	SVC 3	4	5	JRA 6	7	8	Total
1	**STFC**	15	2	3	6	3	18	6		**48**
3	**ESRF**		2	18	3	3		12		**38**
3	**ILL**		2	3	6	3	6	18		**38**
4	**DIAMOND**		2	3	3	3			18	**29**
5	**PSI**		2	18	3	3			12	**38**
6	**DESY**		6	3	3	18			12	**42**
7	**ELETTRA**		2	3	18	3	12			**38**
8	**SOLEIL**		2	3	3	3				**11**
9	**ALBA**		2	3	3	3				**11**
10	**HZB**		2	3	3	3				**14**
11	**CEA/LLB**		2	3	3	3				**11**
	Total	**15**	**26**	**63**	**54**	**51**	**36**	**36**	**41**	**318**

List of Milestones

Mile stone number	Milestone Name	Work package(s) involved	Expected Date	Means of verification
1	Definition of Services	WP3 WP4 WP5	M9	Deliverables 3.1, 4.1 and 5.1 completed as specified.
2	First release of Services	WP3, WP4, WP5, WP6, WP7, WP8	M21	Demonstration of user AAA and data catalogue services as defined in deliverables 3.3 and 4.3
3	Evaluation of Integrated Services	All	M30	Completion of project delivering integrated services as planned.

1.4.4 Graphical presentation of interdependencies



The major dependencies between the technical workpackages in PaNdata

Relies on	Workpackage	Relied upon by
All	**Management**	All
All	**Engagement**	none
External factors only	**User AAA Service**	Data Catalogue and Virtual Labs Services All service releases
Provenance and Preservation JRAs, User AAA Service	**Data Catalogue Service**	All service releases Scalability JRA
User AAA and Data Catalogue Services	**Virtual Laboratories Service**	All services (release2,3,4)
User AAA and Data Catalogue Services	**Provenance JRA**	All services (release 3,4)
User AAA and Data Catalogue Services	**Preservation JRA**	All services (release 4)
Data Catalogue Service	**Scalability JRA**	Data Catalogue Service

1.4.5 Description of significant risks and contingency plans

A risk management process will be established within the overall project management, as detailed in section 2.1. Some risks identified for the management and networking activities are outlined here:

Risk: ***Incompatible policies or standards across facilities***
Type: Internal
Description: Common policies and standards are the focus of the PaNdata Support Action. If these cannot be agreed upon within the time frame planned then the integration of the Services across the facilities may be partial, giving different levels of information from different facilities, and potentially reducing the usefulness of the Services and the impact of the project
Probability: Low – Medium
Impact: Medium-High – reduced exploitation chances
Prevention: Close cooperation between facility managers, early adoption of common policies, appropriate information and dissemination with facilities is already underway within the Support Action. There is some time margin between the scheduled delivery in the support action and the requirement in this project.
Remedies: Standards may be developed which cover only some aspects of the services, or apply to only some of the facilities.

Risk: ***Low acceptance of PaNdata within the scientific community***
Type: Internal and external
Probability: Low – medium
Impact: High – reduced exploitation chances
Prevention:

- Early engagement with other projects and the wider scientific community will be a priority for WP2.
- Service trials and evaluations with end-user base to they can influence design decisions.
- Frequent communication on the added value of PaNdata both within the consortium and outside it.
- Demonstration events and workshops.

Remedies: Analyse and adapt communication and dissemination strategies if necessary.

Risk: ***Insufficient level of collaboration***
Type: Internal and external
Probability: Low-medium
Impact: High: redundant work implying wasted efforts and insufficient visibility and impact of PaNdata in Europe
Prevention: Frequent coordination meetings, staff exchange, close monitoring by the project management board
Remedies: Analyse reasons for insufficient collaboration and revisit the collaboration plan

1.5 Service Activities and associated work plan

The Networking, Service and Research activities in this I3 project are best understood in the context of the project as a whole. For this reason, several tables in this section describe the work plan for the whole project and are repeated verbatim in the sections 1.5 and 1.6 with grey shaded sections to highlight the relevant part. The table below summarises the scope of each subsection.

Section No.	Describes	Scope
1.4.1	Overall strategy of work plan	**Service Activities only**
1.4.2	Timing of the different WPs (GANTT)	Whole project
1.4.3	Work package list	Whole project
1.4.3	Deliverables list	Whole project
1.4.3	Description of each work package	**Service Activities only**
1.4.3	Summary effort table	Whole project
1.4.3	List of milestones	Whole project
1.4.4	Graphical presentation of components and interdependencies (Pert)	Whole project
1.4.5	Risk analysis for service activities	**Service Activities only**

Scope of description of each subsection within this section

1.5.1 Overall Strategy for Service Activities

The overall strategy of the work plan for the whole project is described in Section 1.3. This section describes only those aspects which are specific to the Service Activities

The Service Activities address those elements of the project which relate to the deployment and operation of common integrated services across the participating facilities. There is one workpackage per service and one which will exercise these services in three specific application domains.

The user AAA and data catalogue services build upon existing technology developed elsewhere and so will deliver a first release relatively early in the project. This will form the basis for adaptation and incorporation of the new functionality delivered by the Joint Research Activities through successive quarterly releases later in the project. They will also provide the platform upon which the virtual laboratory services will be built. Although closely linked, the user and data services are considered distinct in order to separate what are logically different concerns and to allow for the potential separate evolution of the authentication and data sharing functionality.

The virtual laboratories service will provide the ultimate demonstration of the utility of the technology provided by PaNdata by illustrating their use in three of the many application domains supported by the participating facilities. It will provide the evidence to support the case for further role out to other application domains beyond the scope of the current project.

Note that for all the Service Activities, the ongoing operation of the service will be integrated into the normal operational activities of the participating facilities. Thus support is only required from the EC for work related to the introduction of the services and the ongoing costs of operating the services will be borne by the facilities themselves. This applies both the running of the services within the project lifespan and beyond. This is reflected in the financial information in the A2 forms as a 50% contribution to the cost of the Service Activities from the partners own resources.

1.5.2 Schedule

Quarters	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Milestones			M1				M2			M3
Network Activities										
Management	D			D				D		D
Engagement	D				D		D		D	
Service Activies										
Users		D		D		D		D		
Data			D		D		D		D	
vLabs		D			D			D		D
Joint Research activities										
Provenance				D		D		D		
Preservation					D		D		D	
Scaleability			D			D				D

Schedule of workpackages for PaNdata ODI

Key.

D mark the quarters in which workpackages have major deliverables.

M1-M3 are the project milestones.

For clarity, dependencies are not marked here but described in the Pert chart later.

The lighter shaded area in the technical workpackages corresponds to periods of time when services are integrated into the normal operations of the facilities.

1.5.1 Detailed Work Description

Workpackage list (with the grey shaded work packages of the networking activities)

Workpackage No.	Work package title	Type of activity	Lead Partner No.	Lead (short name)	Person Months	Start Month	End Month
	Networking Activities						
1	Management	COORD	1	STFC	10	1	30
2	Dissemination	NA	6	DESY	26	1	30
	Total (NAs)				**36**		
	Service Activities						
3	User AAA Service	SVC	2	PSI	63	1	30
4	Data Catalogue Service	SVC	7	ELETTRA	57	4	30
5	Virtual Laboratories	SVC	6	DESY	51	1	30
	Total (SAs)				**171**		
	Joint Research Activities						
6	Provenance	JRA	1	STFC	36	7	30
7	Preservation	JRA	3	ILL	36	10	30
8	Scalability	JRA	4	DIAMOND	36	1	30
	Total (JRAs)				**36**		
	TOTAL (All Activities)				**315**		

Deliverables List (with the grey shaded deliverables of the networking activities)

Del No.	Deliverable Name	WP No.	Nature (R/P/D/O)	Diss. Level	Del. Date
2.1	Project Website	2	O	PU	1
1.1	Project manag't structures, reporting, risk and quality procedures	1	R	CO	3
2.2	Dissemination plan	2	R	PU	3
3.1	Specification of AAA infrastructure	3	R	PU	6
5.1	Specific requirements for the virtual laboratories	5	R	PU	6
4.1	Requirements analysis for common data catalogue	4	R	PU	9
8.1	Definition of pHDF5 capable Nexus implementation	8	P	PU	9
8.2	Evaluation of Parallel filesystems and MPI I/O implementations	8	R	PU	9
1.2	First annual management report	1	R	CO	12
3.2	Pilot deployment of initial AAA service infrastructure	3	P	PU	12
6.1	Model of the data continuum in Photon and Neutron Facilities	6	R	PU	12
2.3	First Open Workshop	2	O	PU	15
4.2	Populated metadata catalogue with data from the virtual laboratories	4	R	PU	15
7.1	Implementation of persistent identifiers for PaNdata datasets	7	D	PU	15
3.3	Production deployment of AAA service infrastructure	3	D	PU	18
5.2	Deployment of Specification of the three virtual laboratories	5	R	PU	18
6.2	Common ontology def'n and def'n of tools to support provenance	6	R	PU	18
2.4	Open Source software distribution procedure	2	R	PU	21
4.3	Deployment of cross-facility metadata searching	4	D	PU	21
7.2	Mechanisms and tools for representation information and archiving	7	R	PU	21
8.3	Implementation of pNexus and MPI I/O on parallel filesystems	8	P	PU	21
8.4	Examination of Distributed parallel filesystem	8	R	PU	21
8.5	Demonstrate capabilities on selected applications	8	D	PU	21
1.3	Second annual management report	1	R	CO	24
3.4	Evaluation of initial AAA service infrastructure	3	R	PU	24
6.3	Tools for building research objects in Photon and Neutron Facilities	6	P	PU	24
2.5	Second Open Workshop	2	O	PU	27
4.4	Benchmark of performance of the metadata catalogue	4	R	PU	27
7.3	Mechanisms and tools for integrity of datasets	7	R	PU	27
8.6	Evaluation of coupling of prototype to multi-core architectures	8	R	PU	27
1.4	Final management report	1	R	CO	30
5.3	Report on the implementation of the three virtual laboratories	5	R	PU	30
6.4	Evaluation report on provenance management	6	R	PU	30
7.4	Report on evaluation of preservation mechanisms	7	R	PU	30

Description of each work package:

Work package no.	3			Start date or starting event:					M1	
Workpackage title	User Catalogue and AAA Service									
Activity Type	SVC									
Part. number	1	2	3	4	5	6	7	8	9	10	11
Part. Short Name	STFC	ESRF	ILL	Diamond	PSI (Lead)	DESY	ELETTRA	Soleil	ALBA	HZB	CEA
Person-months	3	18	3	3	18	3	3	3	3	3	3

Objectives

- To deploy, operate and evaluate a system for pan-European user identification across the participating facilities and implement common processes for the joint maintenance of that system.

Description of work

Methodology:
This task will deploy, operate and evaluate a protocol for introducing a pan-European user identification and single-sign-on system and implement common processes for the joint operation of that system. This is a necessary baseline for enabling seamless cross-facility data access and integration by individual users. It will build on the user policy and user data exchange standards which are being developed by the consortium in the current PaNdata Support Action.

Tasks:

Task1: Consultative process including a survey on existing software components. There should be a gap analysis between AAA requirements and the packages available. This should result in recommendations for technologies to be implemented.

Task 2: Setup of a PaNdata authentication team which includes representatives from the user office and/or IT staff of the partners.

- As the system will in part touch the autonomy of the individual facilities it will be indispensable to have the consensus for the full duration of the project.
- This authentication team will meet in regular intervals.

Task 3: Specify an architecture which meets the requirements of all the participating facilities and builds on the IRUVX "umbrella" concept.

- A very important issue is to ensure that this architecture complies with legal constraints on the transfer of personal information which will have been identified in the current PaNdata Support Action. By basing the system as much as possible on a user-self-service concept, these requirements will be resolved to a large extent. For the definition of the database entries of user affiliations a de-facto standard will be introduced.

- As stated above, the purpose is to extend rather than replace the existing user databases, and in this way to add novel functionalities. Define the adoptions of the local WUO systems for the existing administration, data acquisition and analysis tools. , so that the new functionalities can be used.

Task 4: Implement together with the local user office staff the necessary local modifications (including trust management).

- This should be easily accessible easily by all participants but may not be in the location used for the service activity.
- The system must have an efficient remote management interface.

Task 5: Implement a standard affiliation database which is accessible for update and use by the participating facilities and support the local system managers for migrating to this tool.

- Introduce a central affiliation database according to the PaNdata de-facto standard.
- Provide an interface of the local WUO systems to this standard.
- Organise and support the migration of the local WUOs to this new affiliation database.

Task 6: Deploy the user management system at all participating facilities.

- A major factor will be the integration with the facility's bespoke user administration systems. Here the authentication team will play an important role.
- The deployment will include setting up of an administration authority for the system.
- The final operation will not be contingent on the specific project funding.

Task 7: Evaluate the system within a subset of the collaborating facilities.

- The facilities concerned should have well advanced internal user databases and an implementation of a data storage repository.
- In this pilot period feedback will be collected from actual users on the usability of the system.

Task 8: Operate and report on the AAA trust system for the remainder of the project.

Task 9: Maintain communication with other user authentication systems (through Workpackage 2) and plan future developments of the user management systems to integrate with other systems.

Deliverables and month of delivery

D3.1 : Specification of AAA infrastructure (M6)

D3.2 : Pilot deployment of initial AAA service infrastructure (M12)

D3.3 : Production deployment of AAA service infrastructure (M18)

D3.4 : Evaluation of initial AAA service infrastructure (M24)

Work package no.	4			Start date or starting event:					M4	
Workpackage title	Data Catalogue Service									
Activity Type	SVC									
Part. number	1	2	3	4	5	6	7	8	9	10	11
Part. Short Name	STFC	ESRF	ILL	Diamond	PSI	DESY	ELETTRA (Lead)	Soleil	ALBA	HZB	CEA
Person-months	6	3	6	3	6	3	18	3	3	3	3

Objectives

This workpackage will deploy, operate and evaluate a generic catalogue of scientific data across the participating facilities and promote its integration with other catalogues beyond the project.

Specifically, we will:

1. develop the generic software infrastructure to support the interoperation of facility data catalogues,
2. deploy this software to establish a federated catalogue of data across the partners,
3. provide data services based upon this generic framework which will enable users to deposit, search, visualise, and analyse data across the partners' data repositories,
4. evaluate this service from the perspective of facility users,
5. manage jointly the evolution of this software and the services based upon it,
6. promote the take up of this technology and the services based upon it beyond the project.

Description of work

Methodology:

The metadata catalogue service will build on the policy and data standards developed in the PaNdata Support Action. This work will build on the user AAA services deployed in WP3 and provide a service to the virtual laboratories developed in WP5.

This workpackage will not develop a new metadata catalogue but instead use one of the existing implementations. Inside the community the ICAT from STFC is the most advanced implementation and is already being deployed at 4 of the partner facilities. ICAT is therefore a strong candidate for the baseline for this work. However, we will also analyse and compare with other implementations like the MCA, MCAT Artemis and Fireman.

The technology will build on existing technology already deployed at 4 of the participating facilities. This may need to be adapted however to the current systems at the collaborating institutes.

The following issues will need to be addressed: (1) how to link logical files indexed by metadata to physical files (2) how to query metadata (3) how to authorize user access to metadata (4) what API to propose to programs to access metadata and data.

The first requirement is to analyse the minimum set of keywords required to be included in the metadata catalogue. Building on the existing implementations, and on the output of the PaNdata Support Action workpackage on Integration, an additional set of metadata required by the domains of photon and neutron science may need to be added.

The catalogue will be populated with data from the virtual laboratories (WP5) to demonstrate and test it. It will be possible to fill the data catalogue from existing data archives of the collaborating partners. The work package will demonstrate accessing data distributed over multiple sites via their metadata. The performance and scalability of the metadata catalogue for the virtual laboratories will be evaluated as elaborated in WP5.

Tasks:

Task 4.1. Survey the features of existing implementations of metadata catalogues and compare with metadata, authorisation, performance, and ontological requirements developed in the virtual laboratories (WP5) and the user AAA service specification developed in WP3.

Task 4.2. Serially, deploy the chosen metadata catalogue solution in the legacy context of the collaborating facilities.

Task 4.3. Provide remote API access to the individual catalogues as and integrate to provide a single search capability across the collaborating facilities.

Task 4.4. Evaluate the performance of searching the metadata catalogue and retrieving data.

Deliverables and month of delivery

D4.1. Requirements analysis for common data catalogue (M9)

D4.2. Populated metadata catalogue with data from the virtual laboratories (M15)

D4.3 : Deployment of cross-facility metadata searching (M21)

D4.4. Benchmark of performance of the metadata catalogue (M27)

Work package no.	5			Start date or starting event:					M1	
Workpackage title	Virtual Laboratories									
Activity Type	COORD									
Part. number	1	2	3	4	5	6	7	8	9	10	11
Part. Short Name	STFC	ESRF	ILL	Diamond	PSI	DESY (Lead)	ELETTRA	Soleil	ALBA	HZB	CEA
Person-months	3	3	3	3	6	18	3	3	3	3	3

Objectives

To deploy a set of integrated end-to-end user and data services supporting three specific techniques:

1. Structural 'joint refinement' against X-ray & neutron powder diffraction data
2. Simultaneous analysis of SAXS and SANS data for large scale structures
3. Access to tomography data exemplified through paleontological samples

Description of work

Methodology

Making raw and processed data permanently available to authorised users and the general public world-wide is one of the main aims of PaNdata. Giving scientists access to such data will enable them to complement their private data with published data, limit the duplication of experiments and make the data generally more available to a wider audience who would otherwise not have access to the data e.g. scientists and students who are not users of any of the collaborating facilities.

The three techniques concern data in the fields of diffraction, small angle scattering and tomography applied for example to palaeontology. The first two methods are well-known, the third less well so. Tomography is a technique which provides spectacular 3D images of a wide variety of samples. It typically generates large quantities of data (50 to 100 Gigabytes of processed data). We will take as an example a small subset of tomography users, namely palaeontologists studying samples which are millions of years old in situ. Making new results on hominid and entomological samples results available to a wider public is essential for the paleontological community.

The test cases will :

- demonstrate the integrated use of the services deployed within the project
- do so in the context of commonly-occurring cross-facility analyses of scientific interest
- demonstrate how the services facilitate data analysis or access to data

Tasks:

For each of the three techniques undertake:

- requirements capture in report on existing tools used.
- iterative use of the new tools as they are developed.
- evaluation of the new support (how does the new support compare to the old)

Task 1. Structural 'joint refinement' against X-ray & neutron powder diffraction data.

- Raw data searched for by an authenticated user through the data catalogues.
- Access is authorised and data downloaded from facility archives.
- Relevant analysis software searched for in software database.
- Software downloaded and run locally or at facility.
- Analysis carried out.
- Results (refined structure) and any relevant reduced data uploaded to facility archives.

Task 2. Simultaneous analysis of SAXS (Small Angle X-ray Scattering) and SANS (Small Angle Neutron Scattering) data for large scale structures.

- Raw data searched for by an authenticated user through the data catalogues.
- Access is authorised and data downloaded from facility archives.
- Relevant analysis software searched for in software database.
- Software downloaded and run locally or at facility.
- Analysis carried out.
- Results (modelled structure) and any relevant reduced data uploaded to facility archives.

Task 3. Provide access to tomography data of paleontological samples.

- Setup a public access database for storing tomographic raw and processed data of paleontological data e.g. 2D tomographs and 3D processed images.
- Provide authorised access from multiple institutes to store processed data in the database.
- Enable public access to data in database.
- Implement long term archiving of database.

Deliverables and month of delivery

D5.1: Specific requirements for the virtual laboratories (M6)

D5.2: Deployment of Specification of the three virtual laboratories (incorporating any specific requirements software to support them) (M18)

D5.3: Report on the implementation of the three virtual laboratories (M30)

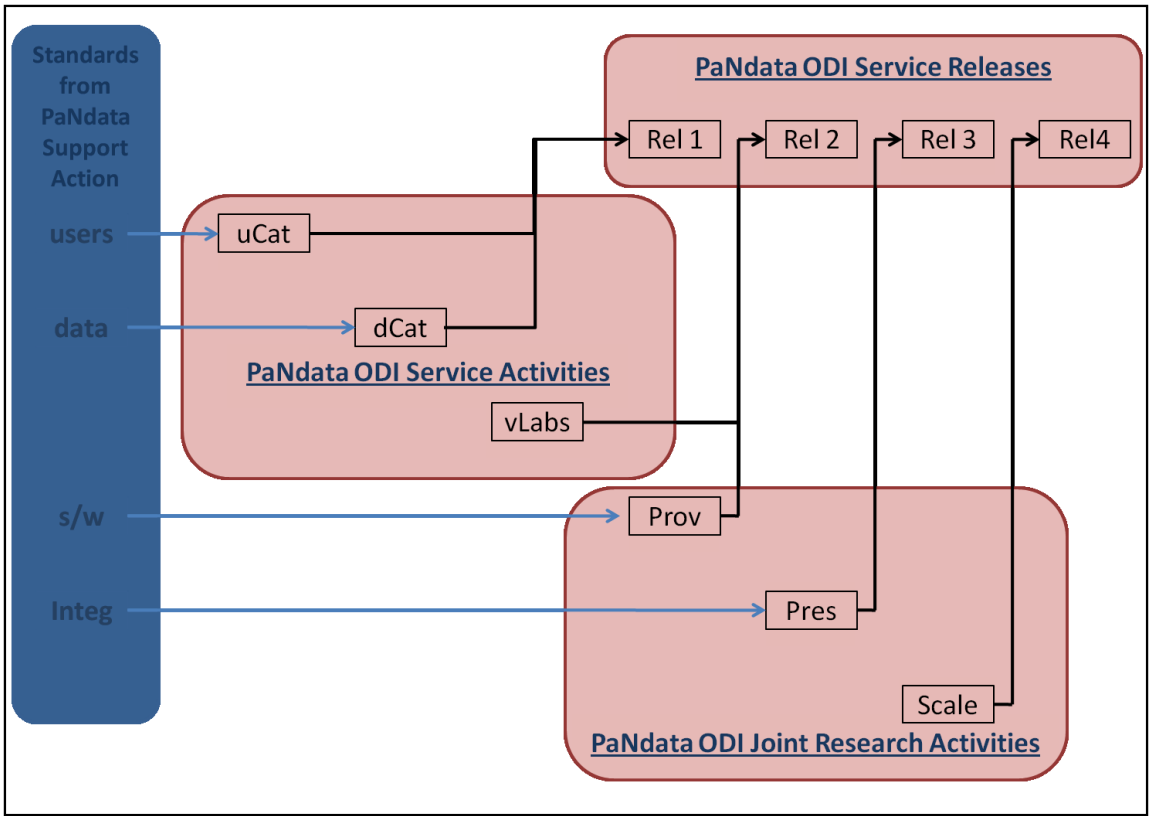
Summary effort table

Partner Number	Short Name	COORD		SVC			JRA			Total
		1	2	3	4	5	6	7	8	
1	**STFC**	15	2	3	6	3	18	6		**48**
3	**ESRF**		2	18	3	3		12		**38**
3	**ILL**		2	3	6	3	6	18		**38**
4	**DIAMOND**		2	3	3	3			18	**29**
5	**PSI**		2	18	3	3			12	**38**
6	**DESY**		6	3	3	18			12	**42**
7	**ELETTRA**		2	3	18	3	12			**38**
8	**SOLEIL**		2	3	3	3				**11**
9	**ALBA**		2	3	3	3				**11**
10	**HZB**		2	3	3	3				**14**
11	**CEA/LLB**		2	3	3	3				**11**
	Total	**15**	**26**	**63**	**54**	**51**	**36**	**36**	**41**	**318**

List of Milestones

Mile stone number	Milestone Name	Work package(s) involved	Expected Date	Means of verification
1	Definition of Services	WP3 WP4 WP5	M9	Deliverables 3.1, 4.1 and 5.1 completed as specified.
2	First release of Services	WP3, WP4, WP5, WP6, WP7, WP8	M21	Demonstration of user AAA and data catalogue services as defined in deliverables 3.3 and 4.3
3	Evaluation of Integrated Services	All	M30	Completion of project delivering integrated services as planned.

1.5.2 Graphical presentation of interdependencies



The major dependencies between the technical workpackages in PaNdata

Relies on	Workpackage	Relied upon by
All	**Management**	All
All	**Engagement**	none
External factors only	**User AAA Service**	Data Catalogue and Virtual Labs Services All service releases
Provenance and Preservation JRAs, User AAA Service	**Data Catalogue Service**	All service releases Scalability JRA
User AAA and Data Catalogue Services	**Virtual Laboratories Service**	All services (release2,3,4)
User AAA and Data Catalogue Services	**Provenance JRA**	All services (release 3,4)
User AAA and Data Catalogue Services	**Preservation JRA**	All services (release 4)
Data Catalogue Service	**Scalability JRA**	Data Catalogue Service

1.5.3 Description of significant risks and contingency plans

A risk management process will be established within the overall project management, as detailed in section 2.1. Some risks identified for the Service Activities are outlined here.

Risk: ***PaNdata infrastructure delayed***
Type: Internal
Description: If the equipment required for implementing the services of WPs 3/4/5 is not ready in due time, then the service activity will be delayed.
Probability: Low – medium
Impact: Medium – implementation of the services in only some of the RIs
Prevention: Strong involvement of the IT responsible of each participating RI, strong coordination between project management board and the IT responsible of each RI.
Remedies: Regular follow up

Risk: ***Code robustness***
Type: Internal
Probability: Medium
Impact: High – may impact the date of production service
Prevention: Use established software development methodology for code quality. Use experienced engineers in software development. Do allow for and insist on extensive debugging. Early start of debugging on specific parts of the code.
Remedies: Reduce the set of functionalities, affect additional resources if appropriate.

Risk: ***Performance below expectations***
Type: Internal
Description: If the performance of one or several services is too low, the user community will not adopt the functionalities.
Probability: Medium
Impact: Medium – adoption of the services in only some of the RIs, or only between some of the RIs.
Prevention: Strong involvement of the IT responsible of each participating RI. Early tests and performance optimisations.
Remedies: Regular follow up

Risk: ***Incompatible pre-existing IT infrastructures across RIs***
Type: Internal
Description: If the existing IT infrastructures across the facilities have different incompatible architectures and systems it may be difficult federating them, thus delaying the service activities.
Probability: Low
Impact: Medium
Prevention: Close collaboration between facility IT managers. Early identification of incompatibilities, mutual visits.
Remedies: Workarounds and specific implementations could be required.

Risk:	***Security systems incompatible across RIs***
Type:	Internal
Description:	If the existing IT infrastructures across the facilities have incompatible security architectures (e.g. firewalls, authentication systems, policies), then federating them may be difficult, thus delaying the service activities.
Probability:	Low
Impact:	Medium
Prevention:	Close collaboration between facility IT managers. Early identification of incompatibilities, mutual visits.
Remedies:	Workarounds could be required.

1.6 Joint Research Activities and associated work plan

The Networking, Service and Research activities in this I3 project are best understood in the context of the project as a whole. For this reason, several tables in this section describe the work plan for the whole project and are repeated verbatim in the sections 1.5 and 1.6 with grey shaded sections to highlight the relevant part. The table below summarises the scope of each subsection.

Section No.	Describes	Scope
1.4.1	Overall strategy of work plan	**JRAs only**
1.4.2	Timing of the different WPs (GANTT)	Whole project
1.4.3	Work package list	Whole project
1.4.3	Deliverables list	Whole project
1.4.3	Description of each work package	**JRAs only**
1.4.3	Summary effort table	Whole project
1.4.3	List of milestones	Whole project
1.4.4	Graphical presentation of components and interdependencies (Pert)	Whole project
1.4.5	Risk analysis for service activities	**JRAs only**

Scope of description of each subsection within this section

1.6.1 Overall Strategy for Joint Research Activities

The overall strategy of the work plan for the whole project is described in Section 1.3. This section describes only those aspects which are specific to the Joint Research Activities.

The Joint Research Activities address those elements of the project which involve the research and development of the technology which underpins the common integrated services across the participating facilities. There is one workpackage per technology required. The technology developed will be incorporated into the later releases of the services.

The Provenance JRA takes the concept of a repository of information about an experiment to a new level. By tracking and logging the data analysis steps it links all the data artefacts across the "data continuum" and thereby allows the tracking of provenance of data "from application to publication".

The Preservation JRA will build upon existing frameworks developed in other initiatives and thus begins from a mature basis. It consists primarily of adapting and modifying these technologies to the current application domains. However some innovative work is expected as detailed in the workpackage description.

The Scalability JRA recognises that handling the advancing "data tsunami" will require effective use of the data transfer and compute infrastructure being developed elsewhere. It will remove some barriers which arise from the essentially serial technology which is currently employed by moving to a more parallel environment.

1.6.2 Schedule

Quarters	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Milestones			M1				M2			M3
Network Activities										
Management	D			D				D		D
Engagement	D				D		D		D	
Service Activies										
Users		D		D			D	D		
Data			D		D		D		D	
vLabs		D			D			D		D
Joint Research activities										
Provenance				D		D		D		
Preservation					D		D		D	
Scaleability			D				D			D

Schedule of workpackages for PaNdata ODI

Key.

D mark the quarters in which workpackages have major deliverables.

M1-M3 are the project milestones.

For clarity, dependencies are not marked here but described in the Pert chart later.

The lighter shaded area in the technical workpackages corresponds to periods of time when services are integrated into the normal operations of the facilities.

1.6.3 Detailed Work Description

Workpackage list (with the grey shaded work packages of the networking activities)

Workpackage No.	Work package title	Type of activity	Lead Partner No.	Lead (short name)	Person Months	Start Month	End Month
	Networking Activities						
1	Management	COORD	1	STFC	10	1	30
2	Dissemination	NA	6	DESY	26	1	30
	Total (NAs)				**36**		
	Service Activities						
3	User AAA Service	SVC	2	PSI	63	1	30
4	Data Catalogue Service	SVC	7	ELETTRA	57	4	30
5	Virtual Laboratories	SVC	6	DESY	51	1	30
	Total (SAs)				**171**		
	Joint Research Activities						
6	Provenance	JRA	1	STFC	36	7	30
7	Preservation	JRA	3	ILL	36	10	30
8	Scalability	JRA	4	DIAMOND	36	1	30
	Total (JRAs)				**36**		
	TOTAL (All Activities)				**315**		

Deliverables List (with the grey shaded deliverables of the networking activities)

Del No.	Deliverable Name	WP No.	Nature (R/P/D/O)	Diss. Level	Del. Date
2.1	Project Website	2	O	PU	1
1.1	Project manag't structures, reporting, risk and quality procedures	1	R	CO	3
2.2	Dissemination plan	2	R	PU	3
3.1	Specification of AAA infrastructure	3	R	PU	6
5.1	Specific requirements for the virtual laboratories	5	R	PU	6
4.1	Requirements analysis for common data catalogue	4	R	PU	9
8.1	Definition of pHDF5 capable Nexus implementation	8	P	PU	9
8.2	Evaluation of Parallel filesystems and MPI I/O implementations	8	R	PU	9
1.2	First annual management report	1	R	CO	12
3.2	Pilot deployment of initial AAA service infrastructure	3	P	PU	12
6.1	Model of the data continuum in Photon and Neutron Facilities	6	R	PU	12
2.3	First Open Workshop	2	O	PU	15
4.2	Populated metadata catalogue with data from the virtual laboratories	4	R	PU	15
7.1	Implementation of persistent identifiers for PaNdata datasets	7	D	PU	15
3.3	Production deployment of AAA service infrastructure	3	D	PU	18
5.2	Deployment of Specification of the three virtual laboratories	5	R	PU	18
6.2	Common ontology def'n and def'n of tools to support provenance	6	R	PU	18
2.4	Open Source software distribution procedure	2	R	PU	21
4.3	Deployment of cross-facility metadata searching	4	D	PU	21
7.2	Mechanisms and tools for representation information and archiving	7	R	PU	21
8.3	Implementation of pNexus and MPI I/O on parallel filesystems	8	P	PU	21
8.4	Examination of Distributed parallel filesystem	8	R	PU	21
8.5	Demonstrate capabilities on selected applications	8	D	PU	21
1.3	Second annual management report	1	R	CO	24
3.4	Evaluation of initial AAA service infrastructure	3	R	PU	24
6.3	Tools for building research objects in Photon and Neutron Facilities	6	P	PU	24
2.5	Second Open Workshop	2	O	PU	27
4.4	Benchmark of performance of the metadata catalogue	4	R	PU	27
7.3	Mechanisms and tools for integrity of datasets	7	R	PU	27
8.6	Evaluation of coupling of prototype to multi-core architectures	8	R	PU	27
1.4	Final management report	1	R	CO	30
5.3	Report on the implementation of the three virtual laboratories	5	R	PU	30
6.4	Evaluation report on provenance management	6	R	PU	30
7.4	Report on evaluation of preservation mechanisms	7	R	PU	30

Description of each work package:

Work package no.	6			Start date or starting event:						M7	
Workpackage title	Provenance										
Activity Type	JRA										
Part. number	1	2	3	4	5	6	7	8	9	10	11
Part. Short Name	STFC (Lead)	ESRF	ILL	Diamond	PSI	DESY	ELETTRA	Soleil	ALBA	HZB	CEA
Person-months	18		6				12				

Objectives

To develop a conceptual framework, which can record and recall the data continuum, and especially the analysis process, and to provide a software infrastructure which implements that model to record analysis steps hence enabling the tracing of the derivation of analysed data outputs.

Description of work

Methodology

Support for raw data and publication is already supported within facilities, so to support the whole data lifecycle, data analysis remains the key link in the chain that transforms experimental results into conclusive scientific output. Different facilities currently provide different levels of support for analysis; we need to consider use cases in all facilities, drawing out best practice and identifying the key drivers and requirements.

From the use cases, a framework identifying a common process and information model will be developed in this workpackage to capture derived data, and to record the analysis process including the analysis software sufficiently for the needs of each facility. This will permit the tracing and logging of the provenance of published data; and to allow access to derived data for secondary analysis, This will be based on current models for capturing raw data, such as the CSMD model underpinning the ICAT suite. In terms of data provenance, the current approach identifies the source provenance of the resultant data product, but it needs to be extended to describe the transformation provenance as well.

In order to use this common information model within particular facilities, it will need to be specialised via domain ontologies identifying facility specific items, such as beamlines, instrument, equipment, experimental methodologies and techniques, and user roles. This workpackage will develop such ontologies to support photon and neutron data management with specific instances for each facility. These ontologies will be supported from within the existing facilities data management tools.

Support for the data continuum needs to be integrated into the working practises of the instrument and end-user scientists. Such tools will need to be as unobtrusive as practicable so that they can easily fitted into the usual working practices, and also be capable of cross-site working, as much analysis is undertaken at the users' home institutions. In this workpackage, suitable tool support will be developed within existing facilities analysis frameworks and APIs, as well as the data management tools such as ICAT. This will exploit and integrate with the common catalogue of facilities software currently under development within the current support action.

In order to exploit the richer framework offered by tracing provenance, existing tool support will

need to be extended to provide views on the data continuum. This would allow for example tracing dependencies on data of a publication, replaying analysis steps, the dependencies on software versions, and citation graphs for data. This work package will extend current user data exploration tools used in facilities to allow the exploration of the data provenance.

Tasks

Task 1: Requirements for Provenance

A survey of the existing approaches and requirements for managing derived data products within the PaNdata facilities. This will consider how data is managed from its raw state through analysis to published data and publication, and how it is merged from different sources, especially from different facilities. Scenarios for capturing and exploiting provenance in PaNdata facilities will be developed.

Task 2: Modelling the data continuum

A model of the appropriate data processes will be developed to support the data continuum process suitable for use across the structural sciences (e.g. biochemistry, chemistry, materials science, earth science, palaeontology etc) supported by the neutron and photon community. This should support an exploratory rather than prescriptive workflow model, allowing logging and reconstruction rather than directing the process.

Task 3: Ontologies for specific instruments/techniques

In order to instantiate the above general model to specific facilities and scientific domains, specialist vocabulary needs to be developed to capture concepts such as: instruments, equipment, techniques, software, samples, people and roles as well as discipline specific vocabulary. In this task we will analyse existing ontologies in this area and consolidate and extend into a common, extensible ontology to support the community.

Task 4: Tool Support for the Data Continuum

Tools to support the data continuum model identified in Task 2 above will be specified, designed and implemented. This will seek to build on and extend existing tools, such as the ICAT information catalogue, and integrate into existing data analysis infrastructures within facilities, via the provision of libraries and APIs which can be used to augment existing tools to capture and utilise data provenance to provide added value services to facility users.

Task 5: Tracing the Data Continuum

Experimental tool support will use the provenance information to provide views on the data continuum for particular purposes, such as tracing dependencies on data of a publication, replaying analysis steps, dependencies on software versions, citation graphs for data. This will enrich the access and interaction with the facilities data assets.

Task 6: Evaluation

An evaluation of tools and methods developed in the JRA will be undertaken to assess the impact and value of provenance management in PaNdata.

Deliverables and month of delivery

D6.1: Model of the data continuum in Photon and Neutron Facilities (M12)

D6.2: Common ontology definition and definition of tools to support the use of provenance for Photon and Neutron Facilities (M18)

D6.3: Tools for building research objects in Photon and Neutron Facilities (M24)

D6.5: Evaluation report on provenance management in Photon and Neutron Facilities (M30)

Work package no.	7		Start date or starting event:						M10	
Workpackage title	Preservation									
Activity Type	JRA									
Part. number	1	2	3	4	5	6	7	8	9	10	11
Part. Short Name	STFC	ESRF	ILL (Lead)	Diamond	PSI	DESY	ELETTRA	Soleil	ALBA	HZB	CEA
Person-months	6	12	18								

Objectives

To incorporate models and tools oriented towards long-term data preservation into the PaNdata infrastructure, focussing on several aspects considered of benefit: an OAIS-based infrastructure; persistent identifiers; and certification of authenticity and integrity

Description of work

Methodology

The approach has a number of parallel and sequential lines of work. The OAIS standard will be applied to the data holdings of the PaNdata facilities to understand the needs for supplementary information oriented towards preservation. This will be supported with metadata schemas and tools that integrate with the scientific process, as for the provenance JRA. One particular aspect of the so-called "representation information" is the tracking of processing done on data sets and publications resulting from them; this relates of course to the work in the provenance work package but with a somewhat different focus.

Complying with the OAIS model will ensure that experimental data that we assume preserved is really in such condition, and it will also help to increase the confidence of our scientific users and their funding bodies in our repositories and help our users to get funded in regards of the upcoming preservation requirements.

A consequence of this is the need to preserve software itself, to allow reprocessing of data for validation and reproduction of results even when the hardware and operating system might have changed. It is not the aim of this project to do research in this area, which is a challenging field in its own right, but to apply the more practical approaches that are emerging.

Within the photon and neutron communities, the NeXus standard data format is being developed and adopted. It is not the aim of this work package to develop the standard; it will be standardised in the Service Activities and work elsewhere. However it is of key importance as an established standard with a long expected lifetime and so will be one of these bases of the preservation work.

A particular need in digital preservation of science data is persistent identifiers. This was identified as one of the elements of the roadmap by the PARSE.Insight project. It is obvious that a prerequisite for a sustainable data infrastructure is the ability to reliably identify particular datasets over time. Issues of granularity, composition and evolution of datasets arise..

Tasks:

Task 1. Baseline and OAIS application

The current situation will be studied in all participating facilities with respect to the objectives listed above. The OAIS standard will be applied and metadata schemas defined for preservation and requirements for integration.

Task 2. Persistent identifiers

A mechanism will be chosen and implemented for creating and referencing persistent identifiers for datasets, allowing long-term linking between raw and derived data and publications.

Task 3. Representation information and archiving

Mechanisms will be set up for creating and maintaining representation information associated with datasets, and for the creation of Archival Information Packages. As far as possible these mechanisms should fit within the normal activities of the scientists and facility staff. This will include software as a kind of representation information, and the need to preserve the software itself.

Task 4. Integrity of datasets

Mechanisms will be established for maintaining and checking integrity of datasets. This is needed both for individual datasets (as preservation actions are performed) and for data holdings as a whole. It includes representation and enforcement of policies on access to data.

Task 5. Evaluation and reporting

Trials will be conducted of the benefits to users of the preservation developments. These will include not only long-term preservation but the ability to retrieve and understand data across disciplines and for the same scientists or team of scientists over a period of time. It will also cover tracking of citations of datasets.

Deliverables and month of delivery

D7.1 Implementation of persistent identifiers for PaNdata datasets (M15)

D7.2 Mechanisms and tools for representation information and archiving (M21)

D7.3 Mechanisms and tools for integrity of datasets(M27)

D7.4 Report on evaluation of preservation mechanisms (M30)

Work package no.	8			Start date or starting event:						M1
Workpackage title	Scalability									
Activity Type	JRA									
Part. number	1	2	3	4	5	6	7	8	9	10	11
Part. Short Name	STFC	ESRF	ILL	Diamond (Lead)	PSI	DESY	ELETTRA	Soleil	ALBA	HZB	CEA
Person-months				18	6	12					

Objectives

To develop a scalable data processing framework combining parallel filesystems with a parallelized standard data format (pNexus pHDF5) to permit applications to make most efficient use of dedicated multi-core environments and to permit simultaneous ingest of data from various sources, while maintaining the possibility for real-time data processing.

Description of work

Methodology:
Several independent developments are enforcing a stronger parallelization of the data processing framework from data acquisition to applications. New detectors (e.g. hybrid pixel array detectors) can produce parallel I/O streams, multi-core environments can process multiple streams simultaneously and parallel (distributed) file systems can cope with the corresponding requirements. To really utilize these highly advanced technologies, a fully parallel, scalable data analyses framework based on an appropriate data format and file system remains to be developed and implemented.

The Hierarchical Data Format (HDF5) has been proposed by EC as a standard for binary data. PaNdata Europe defines HDF5/Nexus within the data policy framework as the standard data format. Nexus is a fully HDF5 compliant extension which just adds a structured, standardized metadata layer. Recently, a parallelized version of HDF5, pHDF5, became available. pHDF5 offers the possibility to tie the different developments - detectors, multi-core CPU/GPU, file systems - together in a fully parallelized data processing framework.

To maintain the particular strength of Nexus providing fully annotated, self-describing and self-containing data, the development of a pHDF5 compliant Nexus API is an important initial step. Independently, the environment most suited for real experimental conditions needs to be investigated. There are a number of different MPI-I/O implementations as well as different parallel file systems on the market, which need to be evaluated, to provide an optimal framework. So far experience with pHDF5/pNexus for typical photon science experiments simply doesn't exist.

Fortunately, pHDF5 capable applications are largely independent of the underlying file system and MPI-implementation. These data analysis applications can hence fully exploit the intrinsic scalability of the data processing framework. This approach can without major modifications extended to distributed environments, for example making use of standard http-protocols to seamlessly analyse data distributed over several different facilities.

To demonstrate and exploit the capabilities of the proposed framework, embedded into the data continuum and authentication infrastructure, selected types of experiments and their applications should be implemented.

We will concentrate on the use cases proposed in the virtual laboratories WP5, namely tomography and crystallography as demonstration show cases for the implementation of pHDF5/pNexus under real experimental conditions.

Tasks:

Task 1: pNexus API.

Develop a pHDF5 compliant Nexus API.

Task 2: Investigate parallel file systems.

Investigate a small number of promising parallel (distributed) file systems with respect to stability, usability, operational costs and efforts support.

Task 3: Investigate implementations on specific file systems

Investigate MPI-I/O implementations and pHDF5/pNexus on an even smaller number of preselected file systems.

Task 4: Coupling of advanced (pre-)processing engines.

Test the capability of the system to cope with multiple parallel data streams. This will contain for example explicit tests feeding a pHDF5-file consisting of a large number of individual images into a multi-core analysis engine.

Task 5: Demonstration.

Implement specific applications in the framework and demonstrate and evaluate the potential of this approach.

Deliverables and month of delivery

D8.1: Definition of pHDF5 capable Nexus implementation (M9) - Software

D8.2: Evaluation of Parallel filesystems and MPI I/O implementations (M9) - Report

D8.3: Implementation of pNexus and MPI I/O on parallel filesystems (M21) - Prototype

D8.5: Examination of Distributed parallel filesystem (M21) - Report

D8.6: Demonstrate capabilities on selected applications (M21) - Demonstrator

D8.7: Evaluation of coupling of prototype to multi-core architectures (M30) - Report

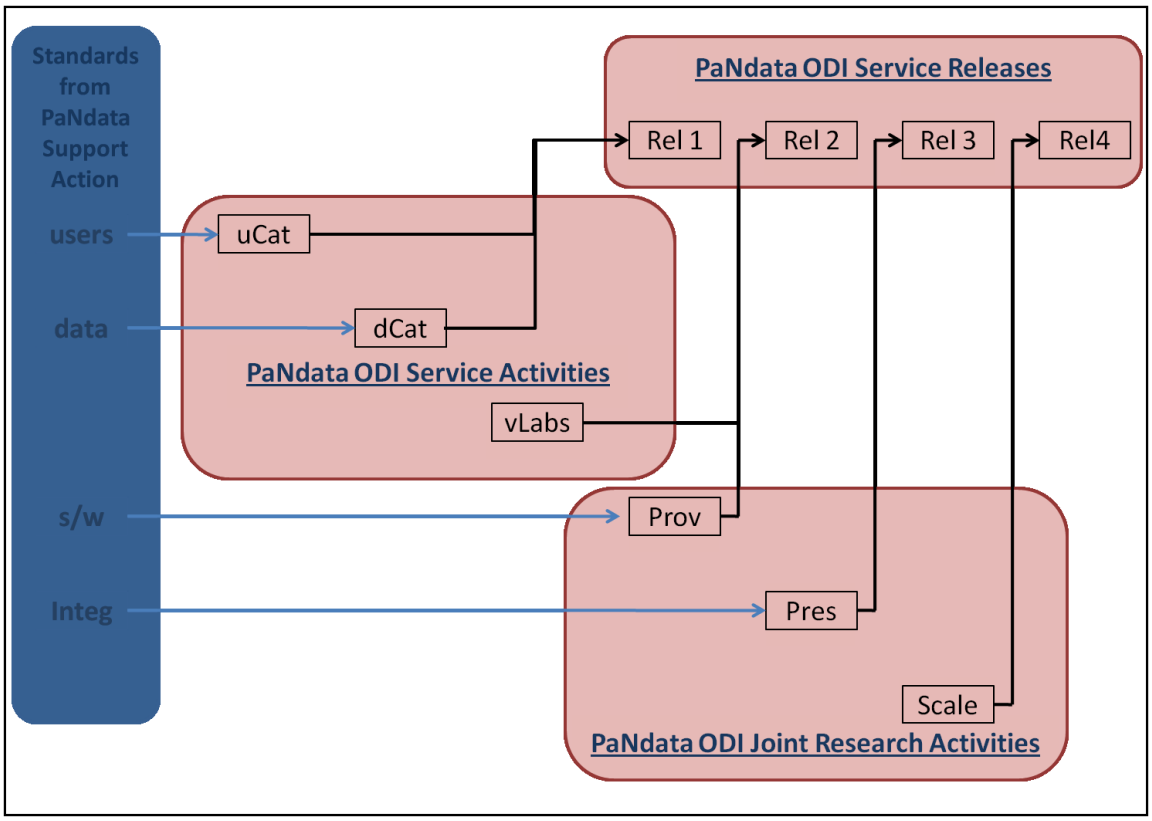
Summary effort table

Partner Number	Short Name	COORD 1	2	SVC 3	4	5	JRA 6	7	8	Total
1	**STFC**	15	2	3	6	3	18	6		**48**
3	**ESRF**		2	18	3	3		12		**38**
3	**ILL**		2	3	6	3	6	18		**38**
4	**DIAMOND**		2	3	3	3			18	**29**
5	**PSI**		2	18	3	3			12	**38**
6	**DESY**		6	3	3	18			12	**42**
7	**ELETTRA**		2	3	18	3	12			**38**
8	**SOLEIL**		2	3	3	3				**11**
9	**ALBA**		2	3	3	3				**11**
10	**HZB**		2	3	3	3				**14**
11	**CEA/LLB**		2	3	3	3				**11**
	Total	**15**	**26**	**63**	**54**	**51**	**36**	**36**	**41**	**318**

List of Milestones

Mile stone number	Milestone Name	Work package(s) involved	Expected Date	Means of verification
1	Definition of Services	WP3 WP4 WP5	M9	Deliverables 3.1, 4.1 and 5.1 completed as specified.
2	First release of Services	WP3, WP4, WP5, WP6, WP7, WP8	M21	Demonstration of user AAA and data catalogue services as defined in deliverables 3.3 and 4.3
3	Evaluation of Integrated Services	All	M30	Completion of project delivering integrated services as planned.

1.6.4 Graphical presentation of interdependencies



The major dependencies between the technical workpackages in PaNdata

Relies on	Workpackage	Relied upon by
All	**Management**	All
All	**Engagement**	none
External factors only	**User AAA Service**	Data Catalogue and Virtual Labs Services All service releases
Provenance and Preservation JRAs, User AAA Service	**Data Catalogue Service**	All service releases Scalability JRA
User AAA and Data Catalogue Services	**Virtual Laboratories Service**	All services (release2,3,4)
User AAA and Data Catalogue Services	**Provenance JRA**	All services (release 3,4)
User AAA and Data Catalogue Services	**Preservation JRA**	All services (release 4)
Data Catalogue Service	**Scalability JRA**	Data Catalogue Service

1.6.5 Description of significant risks and contingency plans

A risk management process will be established within the overall project management, as detailed in section 2.1. Some risks identified for the joint research activities are outlined here.

Risk:	***Incompatible requirements across RIs***
Type:	Internal
Description:	If the requirements across the RIs for the JRAs are too diverging, agreement between the RIs may not be possible.
Probability:	Low
Impact:	High – may lead to blocking situations
Prevention:	Close cooperation between facility managers and the project management board. Since the RIs are working in similar fields, the requirements should be similar.
Remedies:	Standards may be developed which partially cover all aspects of the JRAs and with more detailed specialisations and mappings for a particular facility.

Risk:	***Different software development environments/standards***
Type:	Internal
Description:	If the existing software environments and development cultures in the RIs are very different, it may be difficult making joint software developments.
Probability:	Low – medium
Impact:	Medium – would hamper the exchange and maintenance of code.
Prevention:	Early adoption of common standards
Remedies:	Definition of APIs, concentrating developments more than otherwise necessary

2 IMPLEMENTATION

2.1 Management structure and procedures

2.1.1 Overview of management

The management of the project has the following main objectives:

- to ensure that the project is conducted in accordance with EC rules,
- to reach the objectives of the project within the agreed budget and time scales,
- to co-ordinate the work of the partners and ensure effective communication among them,
- to ensure the quality of the work performed as well as of the deliverables,
- to ensure that appropriate dissemination and outreach is undertaken,
- to ensure that an organisation is set up in order to support the above.

The fulfilment of these objectives is coordinated by Work Package 1 'Management', which will cover those project management activities (administrative, financial, technical co-ordination, IPR, risks…) categorized as management. This work package is placed under the leadership of the Coordinator partner STFC, but defined responsibilities are assigned to all partners.

Budgets will be managed on a per partner basis, rather than per work package.

A ***Consortium Agreement*** will be made between the partners. It will deal with all aspects of the relationships between the organisational bodies stated hereafter, allowing for details such as responsibilities and decision-making procedures, arbitration and project reviewing process. The consortium agreement is being prepared based on standard models.

2.1.2 Project management structure

Given the tight focus of the project, the management structure is relatively simple and depicted in the figure below. It contains the following bodies:

- The **Project Management Board (PMB)** will be chaired by a senior representative from the coordinating facility, and include one representative from each of the partners. The Project Manager will also be a member.
- There will be an **External Advisory Board (EAB)** with external members from the NMI3 (neutron/muon I3), ELISA (synchrotron I3) and e-IRG.
- The **Project Manager (PM)** will be an individual who will manage the operational and reporting activity of the project in collaboration with the Technical Coordination Group. The Project Manager will belong to the coordinating partner, but a different person from the chair of the PMB.
- The **Technical Coordination Group (TCG)** will plan and review the technical progress of the project, and will advise the Project Manager. The group will be chaired by one of the Work Package Coordinators chosen by them all.
- Each work package will have a designated **Work Package Coordinator (WPC)** from the partner identified as the lead for that work package (see tables), responsible for coordination within that work package.

The partners have already established regular methods of contact via e-mail and video conference and these will be continued. Regular face-to-face meetings of project staff will

take place quarterly on a work package basis and short-term staff exchanges are also planned. Formal annual meetings will be attended by board members, work package coordinators and advisory board members.

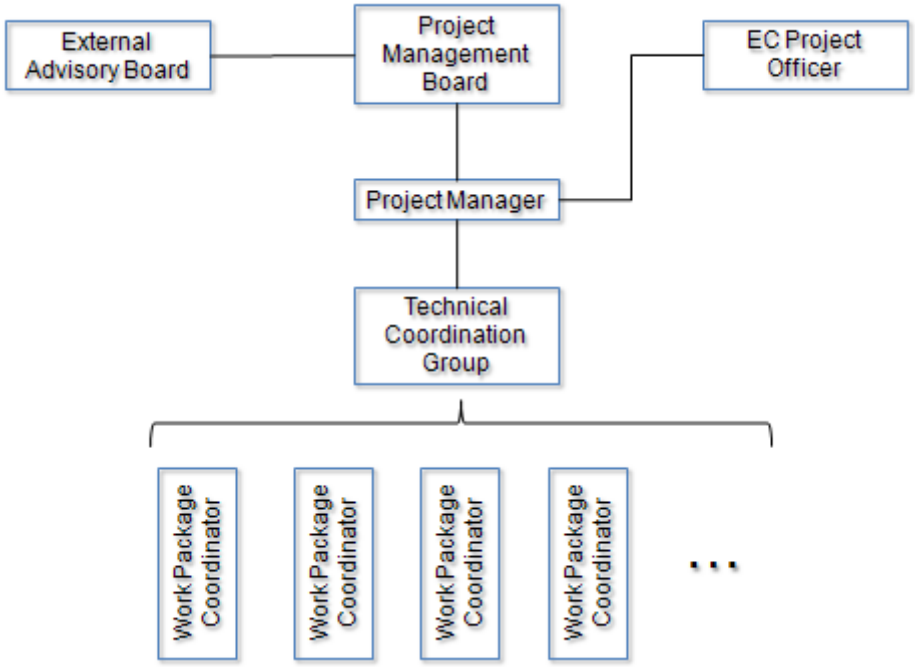


Fig. 2.1: Overview of management structure of the project

2.1.3 Roles and responsibilities

Project Manager. The PM is the interface between the Consortium and the European Commission. The PM is in charge of all administrative and financial matters included in the management work package, particularly:

- ensuring the delivery and the follow-up of administrative and financial documents, including contractual documents, reports, cost statements and funding,
- dealing with questions related to finances, and taking care of the maintenance of the Consortium Agreement and possible contract amendments.

The PM is responsible for the follow up of the deliverables and milestones with help from WP Coordinators.
With respect to the Project Management Board, the PM's duties are:

- to report to the PMB on project progress, especially warning of possible slippage in manpower or resource consumption and planning, so that the PMB can take corrective actions,
- to prepare the agendas of the PMB,
- to monitor the implementation of the decisions of the PMB.

The partner **STFC** which has a thorough experience of EU contracts and is already involved in several consortia of FP6 and FP7 is appointed for this role by the Consortium. **Dr. Juan Bicarregui** from the e-Science Centre, STFC will be appointed project manager for the

duration of the project. His possible replacement is the responsibility of the Project Management Board.

Project Management Board. The Project Management Board is the decision-making body for any strategic issues concerning the operation of the Consortium. It is responsible for the overall control of the Project by its members. In particular, it is the responsibility of the PMB to:

- approve the budget allocation of the EC contribution between the partners, programme of activities and reports,
- decide on contractual changes related to the consortium agreement and EC contract, including in particular changes in the consortium structure and partnership,
- monitor the programme of activities (plans, progress reports, deliverables, funding),
- monitor the performance of the contractors and arbitrating on any conflict arising,
- decide on major IPR issues (publication, licensing, patents and other exploitation of results), subject to the EC Contract and Consortium agreement provisions,
- review upcoming difficulties and risks that may affect the project execution and as such of the implementation of the contingency plan,
- approve all reports and plans to the EC, notably the formal management reports,
- provide any call for and evaluation of new contractors, participants or partners that might be needed to finalize the project objectives,
- liaise with the External Advisory Board and decide on action to take in response to its recommendations.

The PMB consists of at least one representative of each partner, and it is chaired by a senior member of the coordinator partner, **Dr. Robert McGreevy**. The Project Manager will also attend the PMB, but will not have voting rights. A meeting of the PMB will be held at the Project kick-off meeting for validating the activities, the structural methods, the planning and the budget, and then at least 4 times a year.

External Advisory Board. In order for the project to take account of best practice outside the consortium, an External Advisory Board will be established, composed of external members from relevant other projects and initiatives, for example from the NMI3 (neutron/muon I3), ELISA (synchrotron I3) and e-IRG consortia. It will be chaired by one member appointed by the PMB and will give advice on the progress of the project with respect to the wider context. It will also advise on dissemination activities. It will meet on demand, but at least once each year.

Technical Coordination Group. The Technical Coordination group comprises all the Work Package Coordinators, and is chaired by one of them chosen by all its members. The TCG's role is to monitor technical progress, review and propose plans that concern the interaction between work packages, and advise the Project Manager.

Work Package Coordinator. Each work package will have a designated coordinator (an individual person) from the partner organisation identified as responsible for that work package. The WPC will be responsible for scheduling work tasks, allocating resources available, and coordinating the production of deliverables to time and budget. The coordinator will report on progress to the TCG and raise any problems or risks arising from the work package for consideration with other coordinators, the PM and the PMB. The PM and WPCs will consult regularly, through the forum of the TCG, with monthly teleconferences and *ad hoc* discussions as required. The Work Package Coordinators will be chosen by their respective organisations at the start of the project. At the kick-off meeting they will elect the chair of the TCG.

2.1.4 Decision-making process

The ultimate decision making entity of the project is the PMB. However day to day decisions will be made by the PM and WPCs as required. Decisions within the PMB are reached by consensus. In the event that no consensus is reached, decisions will be made by simple majority vote. If this still results in a tie, then the chairman will have the casting vote. Any conflict internal to a work package will be resolved by consensus within the package under the guidance of its coordinator. If the problem could harm normal progress of the project, or have a direct impact on other activities or if it cannot be solved within the activity, the issue will be put to the PMB.

2.1.5 Management of knowledge and IPR

The project outcome will be to a great extent disseminated in form of scientific publications and presentations at conferences or exhibitions. Software and standards arising from the project will be available on an open-source basis and will be disseminated to other large-scale scientific facilities. These activities will be under the co-ordination of the WP2 on Engagement and Dissemination.

The management of knowledge will be carried out according to the usual practice applied by the participants, leaving the maximum access to results to the public. The dissemination and publication of results will meet the contractual requirements in terms of disclosure, and the PMB will check for any IPR issues which may arise.

The management of IPR is an important task of the management work package. The Consortium Agreement will lay down rules for the ownership and protection of knowledge as well as for access rights. In case of disputes, the matter shall be referred to the PMB.

Finally, the WP2 leader will be in charge of collecting and proposing matters referring to the results for dissemination. Once they can be published, an indicator of the productivity of the projects in terms of publications will be provided. A draft plan for use and dissemination of knowledge will be provided as a deliverable of this work package.

2.1.6 Open access

In accordance with the European Commission's Open Access Pilot (see for example ftp://ftp.cordis.europa.eu/pub/fp7/docs/open-access-pilot_en.pdf), the project team will deposit peer-reviewed articles arising from the project into suitable institutional or subject-based repositories, using best efforts to ensure open access to the articles within six months. An example of such a repository already well established within the consortium is STFC's ePubs (http://epubs.stfc.ac.uk).

2.1.7 Risk management and mitigation plan

Risks may have an impact on the project schedule and outcomes, and finally may lead to contractual issues. The project management, coordinated by the PM, shall identify and monitor risks that may have an impact on the project schedule and outcomes and shall take appropriate measures to limit and/or mitigate their effects. The qualitative method applied will be set-up under PM responsibility, applied by all WPCs. It comprises the steps (i) risk identification, (ii) evaluation and ranking, (iii) mitigation and residual risk follow-up. Risk management will be a standing agenda item of all PMB meetings.

Internal risks can result from too ambitious technical objectives and/or unexpected technical difficulty, poor integration of competencies of the participants, deviation from good project management rules, strategy evolutions or defaulting partners.

2.1.8 Quality management

Quality is a key aspect to providing a service to end-users of facilities. Users require a reliable, available, secure, and accurate service to access data and information. The project will establish a quality assurance system, under the responsibility of the PM, and devolved to WPCs for each work package. Each deliverable will be subject to internal review for completeness, accuracy and consistency. Software components will be subject to version control and testing before release. Services will be tested on select user groups to validate their functionality.

2.2 Individual participants

The sections below provide a brief description of each of the participating organisations.

2.2.1 STFC

STFC is the UK public sector research organisation providing access to large scale scientific facilities. It has an expenditure of £500 million p.a. with 2500 staff based at seven locations including the Rutherford Appleton Laboratory (RAL) where this project is centred. Two departments of STFC will be involved in this project.

ISIS is the world's leading pulsed spallation neutron source. It runs 700 experiments per year performed by 1600 users on the 22 instruments. These experiments generate 1TB of data in 700,000 files. All data ever measured at ISIS over twenty years is stored at the Facility, some 2.2 million files in all. ISIS use is predominantly UK but includes most European countries through bilateral agreements and EU funded access. There are nearly 10,000 people registered on the ISIS user database of which 4000 are non-UK EU. The user base is expanding significantly with the arrival of the Second Target Station.

Science & Technology Facilities Council e-Science

e-Science provides the STFC facilities with advanced IT infrastructure including massive data storage, high-end supercomputing, vast network bandwidth, and interoperability with other infrastructure in the UK and internationally. It operates the UK National Grid Service and EGEE Regional Operations for UK and Ireland. It undertakes collaborative IT research at UK, European and global levels. In this project, e-Science will provide overall coordination and provide a bridge to activities such as EGI and eIRG.

Since 2001, e-Science had been developing a common e-Infrastructure supporting a single user experience across the STFC facilities. Much of this is now in place at ISIS and Diamond as well as the STFC Central Laser Facility. Components are also being adopted by ILL, the Australian National Synchrotron and Oakridge National Laboratory in the US.

On ISIS today, experiments instrument computers are closely coupled to data acquisition electronics and the main neutron beam control. Data is produced in ISIS specific RAW format and access is at the instrument level indexed by experiment run numbers. Beyond this data management comprises a series of discrete steps. RAW files are copied to intermediate and long term data stores for preservation. Reduction of RAW files, analysis of intermediate data and generation of data for publication is largely decoupled from the handling of the RAW data. Some connections in the chain between experiment and publication are not currently preserved.

Future data management will focus on development of loosely coupled components with standardised interfaces allowing more flexible interactions between components. The RAW format is being replaced by NeXus. The ICAT metadata catalogue sits at the heart of this new strategy, implementing policy controlling access to files and metadata and using single authentication it allows linking of data from beamline counts through to publications and supports WWW-based searching across facilities.

Dr. Juan Bicarregui is Head of the e-Science Applications Support Division which provides e-infrastructure technology for the STFC facilities and National and European data preservation initiatives such as the UK Digital Curation Centre and the Alliance Permanent Access and the PARSE-Insight and SOAP Support Actions. He has extensive experience in European projects including previously coordinating an FP5 ESPRIT project.

Prof. Robert McGreevy is Head of the ISIS Instrumentation, Diffraction and Muons Division. He has considerable experience of project coordination, for example, the Integrated Infrastructure Initiative for Neutron Scattering and Muon Spectroscopy, the ISIS EU-TS2 Infrastructure Construction project, and of the Neutron I3-Network. **Dr. Brian Matthews** is leader of the Information Management Group in e-Science. He led the development of the CSMD metadata model behind ICAT and the STFC publications archive.

2.2.2 ESRF

The European Synchrotron Radiation Facility is a third generation synchrotron light source, jointly funded by 19 European countries. It operates 40 experimental stations in parallel, serving over 3500 scientific users per year. At the ESRF, physicists work side-by-side with chemists, materials scientists, biologists etc., and industrial applications are growing, notably in the fields of pharmaceuticals, petrochemicals and microelectronics. It is the largest and most diversified laboratory in Europe for X-ray science, and plays a central role in Europe for synchrotron radiation. The ESRF is currently engaging in a development programme for the next 10 years referred to as the Upgrade Programme. International collaborations will be paramount for the success of the ESRF Upgrade Programme, and cover many scientific disciplines including instrumentation and computing developments. ESRF provides the computing infrastructure to record and store raw data over a short period of time and also provides access to computing clusters and appropriate software to analyse the data. The ESRF will witness a dramatic increase in data production due to new detectors, novel experimental methods, and a more efficient use of the experimental stations. The Upgrade Programme will push a significant part of the ESRF beamlines to unprecedented performances and will further increase the data production from currently 1.5 TB/day by possibly three orders of magnitude in ten years from now.

The ESRF has a long track record of successful international collaborations in many different fields of science and technology (SPINE, BIOXHIT, eDNA, X-TIP, SAXIER, TOTALCRYST, etc.). Three international projects are of direct relevance to PaNdata – the international TANGO control system collaboration, ISPyB, and SMIS. The TANGO control system was initially developed for the control of the accelerator complex and the beamlines at ESRF and has been adopted by SOLEIL, ELETTRA, ALBA, and DESY. It shows that five of the PaNdata partners are already working together in software developments of common interest. ISPyB is part of the European funded project BIOXHIT for managing protein crystallography experiments. In its current state, it manages the experiment metadata and data curation for protein crystallography. The SMIS project is the ESRF's database for handling users and experiments.

Andy Götz worked on beamline control, data acquisition, on-line data analysis and Grid technology. He has recently been nominated as the Head of the Software group within the Instrumentation Development Division. He is internationally known for his contributions in control system developments, is member of the NeXus advisory committee and of the ICALEPCS ISAC. He has degrees in computer science and radio astronomy.

Dominique Porte is the group leader of the Management Information System group at the ESRF. He has considerable experience with the design of database systems and is the chief architect of the ESRF proposal submission system (SMIS).

Rudolf Dimper is the Head of the ESRF Computing Services Division. This position entails defining the computing policy of the laboratory, managing the associated resources, and representing the laboratory in computing matters on an international level. He has a degree in chemical engineering.

Manuel Rodriguez-Castellano is the Head of the Industrial and Commercial Unit and Head of the DG's Office. Under his leadership, the Industrial and Commercial Unit deals with all formal aspects of European collaboration contracts. He is a lawyer and has an MBA degree.

2.2.3 ILL

The Institut Laue-Langevin (ILL), founded in 1967, is the European research centre operating the most intense slow neutron source in the world. It is owned and operated by its three founding countries – France, Germany and the United Kingdom – whose grants to the Institute's budget are enhanced by 11 other European partners. ILL is a major player in the European neutron community networks, ENSA and FP7 (NMI3, ESFRI), working with the European Commission to establish and support R&D programs on neutron technology, networks of excellence and workshops. It is also a member of the EIROforum collaboration between seven of Europe's foremost scientific research organizations.

The ILL's mission is to provide the international scientific community with a unique flow of neutrons and a matching suite of experimental facilities (some 40 instruments) for research in fields as varied as solid-state physics, material science, chemistry, biology, nuclear physics and engineering. The Institute is a centre of excellence and a world leader in neutron science and techniques. Every year about 2000 scientists visit the ILL from over 1000 laboratories in 45 different countries across the world to perform as many as 750 experiments per year.

The ILL has a fully-functional computing environment that covers all aspects of experiment and data management; most of the tools have been running for many years and continue to evolve, but they are not shared with any other RI. All neutron data since the start of the ILL is stored. Data collected since 1995 is easily available using Internet Data Access (IDA). This service will be replaced in the near future by a new catalogue based on the iCAT project, enhancing functionality and compatibility with other RI's. On new instruments with very large detectors (BRISP and IN5), the traditional ILL data format has been replaced with a NeXus format, which will be rolled-out to all instruments. Standardised file formats based on NeXus, which are already compatible with the main data treatment codes at ILL, will facilitate the inter-operability of data and software between RI's.

The Scientific Coordination Office (SCO) has a data base of users and the "ILL Visitors Club" is a user portal which constitutes a web-based interface to the SCO Oracle database. The data base (and the information stored in it) is shared by different services at the ILL through different web-interfaces and search programs adapted to their needs. The ILL Visitors Club includes the electronic proposal and experimental reports submission procedures and makes available additional services on the web, such as instrument schedules, user satisfaction forms and information for scientific committees.

Jean-François Perrin is the head of the ILL IT department; his role is to manage the team responsible for the maintenance and improvement of the general aspect of informatics and telecommunication.

Mark Johnson is the head of the Computing for Science group, which is responsible for data analysis software, with input on related issues like data formats, and instrument and sample simulations

2.2.4 Diamond

Diamond Light Source (http://www.diamond.ac.uk/) is a new 3rd generation synchrotron light facility. It became operational in January 2007 and is the largest scientific facility to be funded in the UK for over 40 years. The UK Government, through STFC, and the Wellcome Trust have invested £380M to construct Diamond and its first 22 beamlines of which currently 13 are operational with the remaining 9 entering service in the next few months. Diamond will ultimately host as many as 40 beamlines, supporting the life, physical and environmental sciences.

Diamond's X-rays can help determine the structure of viruses and proteins, important information for the development of new drugs to fight everything from flu to HIV and cancer. The X-rays can penetrate deep into steel and help identify stresses and strain within real engineering components such as turbine blades. They can help improve process for the manufacture of plastics and foods by allowing scientists to observe changing conditions, as well as helping scientists develop smaller magnetic recording materials - important for data storage in computers. The active user population is growing rapidly and will soon exceed 1000 users drawn from the UK, the rest of Europe and indeed the rest of the world.

The Diamond e-Infrastructure supports an integrated data pipeline comprising several shared components. The same configurable Java based Generic Data Acquisition (GDA) system is used across the beamlines. The low level control system is the widely used EPICS system which provides a stable and reliable means for hardware control. Diamond has worked closely with ISIS, and the STFC Central Laser Facility, e-Science and the central site services to implement a cross site user authentication system. Diamond has collaborated with the ESRF and ISIS to implement Web based user administration (DUODESK) and proposal submission (DUO) applications.

The DUODESK application is integrated with most aspects of user operation ranging from accommodation and subsistence through to system authentication, authorization and metadata retrieval.

Diamond is currently working with STFC e-Science and ISIS to provide an externally available data storage repository based on the Storage Repository Broker (SRB) with the ICAT database.

Dr. Bill Pulford. Bill Pulford is currently head of the Data Acquisition and Scientific Computing group at the Diamond Light Source. He has performed similar roles first at the ISIS neutron facility and later at the European Synchrotron Radiation Facility. He has very extensive experience at most aspects of data acquisition with both X-Rays and Neutrons. He was one of the earliest instigators of data management at ISIS and is currently a prime mover in a Single Sign On (SSO) project across UK research facilities.

Dr. Alun Ashton. As a member of the Scientific Computing and Data Acquisition Group at Diamond Light Source, Alun Ashton is responsible for coordinating data analysis activities across all Diamond beamlines. In addition to driving and leading the scientific requirements for internal diamond usage of eScience infrastructure, he has extensive experience of leading roles or working in scientific collaborations such as CCP4 (Collaborative Computational Project Number 4), the DNA project (a project on Automated Data Collection and Processing at Synchrotron Beamlines), Protein Information Management System (PIMS) Project, and has participated in a number of European initiatives such as Autostruct, Maxinf (FP5) and BioXHIT (FP6).

2.2.5 PSI

Within the Swiss research and education landscape, PSI (Paul Scherrer Institut, http://www.psi.ch), plays a special role as a user lab, developing and operating large, complex research facilities. The two large-scale PSI facilities, the Swiss Light Source (SLS) for photon science and the Neutron Spallation Source (SINQ), are responsible for more than 3,000 user visits per year, about half of them international. During the 20 year history of PSI, nearly 20,000 external researchers have performed experiments in the fields of physics, chemistry, biology, material sciences, energy technology, environmental science and medical technology. The Swiss Light Source (SLS) is a third-generation synchrotron light source. With an energy of 2.4 GeV, it provides photon beams of high brightness for research in materials science, biology and chemistry with 16 beamlines in user operation (2009) and 18 as final number. The Spallation Neutron Source (SINQ) is a continuous source - the first of its kind in the world - with a flux of about 1014 n/cm2/s. Besides thermal and cold neutrons for materials research and the investigation of biological substances. The PSI X-ray Free Electron Laser (SwissFEL) is a new development in laser and accelerator-technology. Innovative concepts in accelerator design will limit the overall length of the facility to 800 m. With three branches, it will cover the wavelength range from 10 nm (124 eV) to 0.1 nm (12.4 keV). The SwissFEL should go into operation in 2015. Since decades, PSI researchers are engaged in collaborations for experiments at the PSI facilities, at CERN, ESRF and other large facilities. Initially started as a spin-off of the participation in the CMS detector at LHC, the PSI detector group has developed large-area 1D and 2D photon detectors (Mythen and Pilatus).

The current data acquisition and data storage environment is heterogeneous: various machine and beamline operational parameters are provided by the facilities but there is no standard for recording metadata. SINQ uses the in house program SICS for data acquisition. Most SINQ instruments already store their raw data in the NeXus format. All SINQ data files ever measured are held on an AFS file system and are visible to everyone. Data acquisition at SLS is based on the EPICS system. Data measured at SLS is stored on central storage for two months only. Users are supposed to take their data home on portable storage devices. There is only very limited support for data analysis at SLS.

Stephan Egli is the head of the PSI Information Technology division. He has long term experience as the software WPL of a large HEP collaboration and experience with the needs of researchers in particular in the area of efficient mass data handling. He has a degree in high energy physics.

Derek Feichtinger is head of PSI's Scientific Computing section. He has been involved in the LHC Grid and European Grid projects since 2002 and in building up and running the Swiss LHC Grid Tier-2 centre. He has a degree in Chemistry.

Mark Koennecke is responsible for data acquisition and software for the spallation neutron source SINQ. He is also a long-time member of the NeXus International Advisory Committee and one of the co-inventors of the NeXus data format. He has a degree in materials science.

Heinz J. Weyer has led in the past the group that developed the Digital User Office in use at many European facilities; he was scientific WPL of the SLS. Currently he is involved in several FP7 programs, mostly in connection with IT projects. He has a degree in high energy physics.

2.2.6 DESY

DESY (http://www.desy.de) has a long history in High Energy Physics (HEP) and Synchrotron radiation. DESY runs a Tier-1 centre for the LHC project and has proven expertise in the management and storage of very large data volumes. DESY jointly provides the major software framework (dCache) for large scale and secure data storage. DESY is currently establishing the infrastructure for long term archival and management of the data and metadata from all photon science experiments on site, enabling remote access to data as well as dedicated compute resources, the PaNdata data policy framework being a crucial element for this effort.

While HEP remains an important pillar at DESY, the main focus has clearly shifted towards photon science. DESY is nowadays operating two dedicated synchrotron sources (Doris and Petra III) as well as a free electron laser for the VUV and soft X-ray wavelength regime. (FLASH). Although Petra III, the most brilliant synchrotron source world wide, became operational only very recently, an extension of Petra III to host additional instruments is already in planning phase. The construction of the European XFEL (www.xfel.eu), is progressing well and construction of a second FLASH facility will start soon, accompanied by the foundation of a Center for Free Electron Lasers (CFEL) as well as a Center for Structural and System Biology (CSSB). In parallel, detector development is rapidly progressing, which will allow to obtain diffraction images at a sub-millisecond timescale to cope with the unique time structure of the European XFEL laser light.

DESY, in close co-operation with the Max-Planck Society (MPG), the European Molecular Biology Laboratory (EMBL) and the Helmholtz Centre Geesthacht, which operate several instruments at DESY lightsources, supports several thousand users per year performing photon science experiments, ranging from material sciences to tomography of biological samples. To fully exploit the scientific opportunities at the different light as well as neutron sources, the standardization of experiment descriptions, data formats and policies across facilities is a crucial element. Based on this, implementation of new technologies fully exploiting the capacities of parallel software and dedicated multi-core architectures will become feasible, thereby creating a scalable infrastructure for new analysis and data flow frameworks.

Therefore, DESY will within this project mainly focus on development and implementation of pHDF5 capable applications on suitable parallel filesystems, metadata and data transfer. Since DESY is one of the lead partners in PNI-HDRI, CRISP, EuroFEL and the European X-FEL, dissemination to support compatibility of developments will be the other main task.

Thorsten Kracht is the head of the experiment control group of the photon science department at DESY. His group supports the synchrotron radiation experiments in various fields: online computing, digital user office, electronics, web services and computer administration. He has a degree in physics.

Frank Schluenzen is a member of IT-Department at DESY, coordinator of PNI-HDRI and actively involved in other projects like EuroFEL, WissGrid or CRISP. Formerly working on ribosome crystal structures, he has a strong background in Synchrotron Radiation experimental and computational techniques. He has a degree in Physics.

2.2.7 ELETTRA

ELETTRA (http://www.elettra.trieste.it) is a national laboratory located in the outskirts of **Trieste (Italy). Its mandate is a scientific service to the Italian and international research** communities, based on the development and open use of light produced by synchrotron and Free Electron Lasers (FEL) sources. The ELETTRA infrastructure consists of a State of the art (2-2.4) GeV electron storage ring and about 30 synchrotron radiation beam lines with 13 insertion devices. ELETTRA covers the needs of a wide variety of experimental techniques and scientific fields, including photoemission and spectromicroscopy, macromolecular crystallography, low-angle scattering, dichroic absorption spectroscopy, and x-ray imaging serving the communities of materials science, surface science, solid-state chemistry, atomic and molecular physics, structural biology, and medicine.

ELETTRA is now building a new light source called FERMI@Elettra which is a single-pass FEL user-facility covering the wavelength range from 100 nm (12 eV) to 10 nm (124 eV). This new research frontier of ultra-fast VUV and X-ray science drives the development of a novel source for the generation of femtosecond pulses.

At ELETTRA each beamline has its own acquisition system based on different platforms (java, LabVIEW, IDL, python, etc.). To offer a uniform environment to the users where they can operate and store data, ELETTRA has developed the Virtual Collaboratory Room (VCR) that, among other things, allows users to remotely collaborate and operate the instrumentation. This system is a web portal where the user can find all the necessary tools and applications; i.e. the acquisition application, the data storage, the computation and analysis, the access of remote devices and almost everything necessary for the completion of the experiment. The system implements an Automatic Authentication and Authorization (AAA) based on the credential managed by the Virtual Unified Office (VUO). The VUO web application handles the complete workflow of the proposals' submission, evaluations, and scheduling. The system can provide administrational and logistical support i.e. accommodation, subsistence, access to the ELETTRA site.

The participating team has gained experience in Grids by participating in a set of FP6 EU founded projects like EGEE-II (Enabling Grids for E-SciencE), GRIDCC (Grid Enabled Instrumentation with Distributed Control and Computation) and EUROTeV. GRIDCC introduced the concept of Grid enabled instrument and sensor which is extremely important for industrial applications. Experience gained in FP6 projects is being capitalised as ELETTRA is also participating in the DORII project (Deployment of Remote Instrumentation Infrastructure) and in the Italian Grid Infrastructure. ELETTRA hosts a Grid Virtual Organization (including all the necessary VO-wide elements like VOMS, WMS, BDII, LB, LFC, etc.) and provides resources for several VOs. The current effort is on porting many legacy applications to a Grid computing paradigm in an effort to satisfy demanding computational needs (e.g. tomography reconstruction).

Dr. Roberto Pugliese is a research WPL at Sincrotrone Trieste S.C.p.A. leading the Scientific Computing Group. Since October 2002 he is also Professor of E-Commerce at the University of Udine. His research interests include Web Based Virtual Collaborations and Grid technologies. Roberto Pugliese was the technical WPL of the GRIDCC project and is currently coordinating the Applications workpackage of the DORII project.

Dr. Roberto Borghes is a senior technologist at Sincrotrone Trieste S.C.p.A. where he is a member of the Scientific Computing Group. He is an expert of data acquisition, data treatment and beamline automation.

2.2.8 Soleil

The **Synchrotron SOLEIL** (http://www.synchrotron-soleil.fr) is a 3rd generation synchrotron radiation facility in operation since 2007. In 2009, 1,719 users have performed 348 experiments on the 14 first open beamlines. Currently, SOLEIL is delivering photons to 21 beamlines with a current of 400 mA in top-up mode: 17 beamlines are open to users and 4 under commissioning. In addition, new challenging beamlines are under construction or under design, while SOLEIL is developing technical platforms as the IPANEMA one for Cultural Heritage research. More than 2,000 users from France, Europe and other countries are expected per year to perform experiments in various fields as surface and material science, environmental and earth science, very dilute species and biology.

Responsibility for operating the SOLEIL facility is under the charge of its two shareholders, the CNRS (72%) and the CEA (28%). SOLEIL is involved in bilateral partnerships with more than 12 Universities and Research Institutes and about 30 collaborative projects for ANR and the European Research Programmes have been successfully supported. On the Computing and Controls side, a great effort has been made very early to standardise hardware and software, keeping in mind developments reusability and easy maintenance. The data acquisition system of each Beamline is based on the **TANGO** system, also used for the Machine control. All beamlines can automatically generate data in the **NeXus** standard format, ensuring easier data management and contributing to future interoperability with other research facilities. NeXus files are stored via the storage infrastructure managed with the Active Circle software, handling data availability, data replication on disks and tapes, lifecycle management. Data are accessible from the beamlines as well as from any office in the buildings, with security based on LDAP authentication. A remote access search and data retrieval system, TWIST, allows users to perform complex queries to find pertinent data and to download all or parts of a NeXus file. Data post-processing is handled either on local PCs, or on a beamline compute cluster (if required for experiment control), or on a central HPC system.

Brigitte Gagey is the head of SOLEIL IT Division, defining the computing policy and managing all resources involved in Electronics, Controls and Computing. She has a long time experience at CEA on computing services for the TORE SUPRA Tokamak facility. She holds a degree in plasma physics.

Alain Buteau is the Data Acquisition and Control software group leader, covering from low-level software interfacing electronics and equipments up to Graphical User Interfaces, for Machine and Beamlines needs. Previously, he was in charge of computing and BL controls resources of the LLB neutron facility at CEA.

Philippe Pierrot is the Systems and Network group leader, taking care of all resources pertaining to Office Automation, High Performance Computing, Scientific Data Storage, as well as the network infrastructure for the whole facility.

Jean-Marie Rochat is the Database Management group leader, handling all tasks related to database design and operation, including the Experiment Data Management system. Previously, he was in charge of the LURE management information and proposals systems.

Pascale Prigent is the Instrumentation and Coordination group leader in the Experimental Division. One team of the group is responsible for the coordination and development of software for specific experiments and data analysis. She holds a degree in plasma physics.

2.2.9 ALBA

ALBA is a third generation synchrotron facility near Barcelona, Spain to be constructed and exploited by the consortium CELLS financed equally by Spain and Catalonia. It will include a 3 GeV low emittance storage ring which will feed an intense photon beam to a number of beamlines dedicated to basic and applied research. The accelerator complex will consist in a 100 MeV Linear Accelerator and a Booster that will ramp the electron beam energy up to the nominal energy of 3 GeV. The maximum operational design current is 400 mA and it will be operated in top up mode.

In the first phase, an ensemble of seven beamlines will be operational in 2010. In the subsequent Phases, more beamlines are expected to be built. Phase I beamlines are state of the art in terms of optics and instrumentation. They are as follows: 1) Non Crystalline Diffraction beamline (NCD) for SAXS and WAXS experiments, 2) Macromolecular Crystallography (XALOC), 3) Photoemission (CIRCE), 4) X-ray absorption spectroscopy (XAS), 5) High Resolution Powder Diffraction (MSPD), 6) X ray Circular Magnetic Dichroism (XMCD) and 7) X ray microscopy (MISTRAL). These initial beamlines are designed to cover a wide range of fields such us material science, nanotechnology, medicine, physics, chemistry.

As a new facility, ALBA is starting to participate in European projects and is actively seeking to support not only the Spanish but also the European scientific community. For example ALBA is participating in the project named PRE-XFEL (FP7-211604) to carry out the preparatory activities for the implementation of the European X-ray Free Electron Laser Facility.

The Linac and Linac-to-Booster transfer line have been commissioned and the booster will start its first commissioning phase in December. The installation of the storage ring is already far advanced and according to schedule. Hutches for many beamlines are already built and first optical elements installed. First users are expected at the end of 2010 with routine user operation in 2011.

Computing and Control is largely centralised in one division. The division takes care of the infrastructure (e.g. cabling and racks), electronic support and development, control software, the personal and machine safety system, scientific software, machine timing, systems (central storage, central and individual computing resources, and the network), management information services, the WEB, and the ERP. The accelerator control system is done with Tango, Sardana Pool, and Tau based on C++ and Python for the software and on PCI, cPCI, and PLCs for the hardware. ALBA is actively participating in the TANGO collaboration and is leading the development in the new generic data acquisition system Sardana in collaboration with the ESRF and DESY. The main purpose of the division is to support its internal customers and the future users of the synchrotron.

Having already developed a broad basis for standardization, ALBA is very interested to actively participate in software and hardware developments, common policies and discussions, and sharing of resources with other labs.

Joachim Metge is the Head of the System Section at ALBA which is responsible for providing the hardware resources for all computing needs including network, printing, user computers and central computing facilities. He holds a degree in physics.

Jörg Klora is the Head of the Computing and Control Division and member of the ALBA management board. He holds a degree in physics.

2.2.10 Helmholtz Zentrum Berlin für Materialien und Energie

The Helmholtz Zentrum Berlin (HZB) has emerged in the beginning of 2009 from the merger of BESSY and the Hahn-Meitner Institute. The new centre thus operates two large scale facilities for the investigation of structure and function of matter: the research reactor BER II, for experiments with neutrons, and the electron storage ring facility BESSY II for the production of synchrotron radiation. The HZB also operates the Metrology Light Source, a dedicated storage ring for the German National Metrology Institute PTB (Physikalisch-Technische-Bundesanstalt).

The storage ring BESSY II in Adlershof is at present Germany's largest third generation synchrotron radiation source. BESSY II emits extremely brilliant photon pulses ranging from the long wave terahertz region to hard X rays. The 46 beamlines at the undulator, wiggler, and dipole sources offer users a many-faceted choice of experimental stations. The combination of brilliance and photon pulses makes BESSY II the ideal microscope for space and time, allowing resolutions down to femtoseconds and picometres.

The research reactor BER II delivers neutron beams for a wide range of scientific investigations, in particular for materials sciences. Both thermal and cold neutrons are generated and used for experiments on a total of 24 measuring stations. The HZB offers highly specialised sample environments, allowing for such experiments to take place in high magnetic fields and a wide range of temperatures and pressure.

The HZB aims at strengthening the complementary use of photons and neutrons for basic and applied scientific research. The centre's activities are mainly geared towards a service for an international scientific research: Every year the HZB user service arranges access to its facilities for some 2,500 external scientists (from 35 countries to date). About 100 doctoral candidates from the neighbouring universities are involved in research and training at HZB. The HZB also has extensive experience in scientific collaboration, as many beamlines and experimental stations have been build in collaboration with external research groups. There is an ongoing commitment to develop hardware and software in collaboration with other institutions for the broader scientific community. To date the HZB cooperates with more than 400 partners at German and international universities, research institutions and companies.

Currently many activities focus on merging the technical and scientific support of the centre, in order to provide a more homogeneous and more effective work environment for its users. To this end the HZB also welcomes and participates in European initiatives, as for example on joint user-portals and cross-site AAA-schemes within the ESRFUP and EuroFEL work packages. With respect to its control systems, BESSY has always been a major contributor to the EPICS project and will continue to do so under the HZB banner.

Dr. Dietmar Herrendörfer is deputy head of the HZB's experiment IT department, dealing with beamline control, data acquisition and remote access issues. As a physicist within the IT department, he is also coordinating scientific requirements with the technical focus of the HZB's IT services.

Matthias Muth is head of the HZB's network, storage and server department and responsible for HZB's IT policies and operations, in particular dealing with networking and data storage. He has considerable experience in the design and implementation of high availability clusters and data storage.

2.2.11 CEA/LLB

The French Atomic Energy Commission (CEA: Commissariat à l'énergie atomique) is a public body leader in research, development and innovation. The CEA mission statement has two main objectives: To become the leading technological research organization in Europe and to ensure that the nuclear deterrent remains effective in the future. The CEA is active in three main fields:

- Energy,
- Information and health technologies,
- Defence and national security.

In each of these fields, the CEA maintains a cross-disciplinary culture of engineers and researchers, building on the synergies between fundamental and technological research. In 2008, the total CEA workforce consisted of ~15 000 employees (52 % of whom were in management grades).

Within CEA, the Léon Brillouin Laboratory (LLB) is the National Laboratory of neutron scattering, serving science and industry. The LLB uses the neutrons produced by Orphée, a fission reactor of 14 MW of power. The LLB-Orphée facility is supported jointly by the CEA and the National Centre for Scientific Research (CNRS: Centre National de la Recherche Scientifique). The CEA operates the reactor Orphée located at the Centre d'Etudes de Saclay, since 1980. The LLB gathers the scientists who operate the neutron scattering spectrometers installed around the reactor Orphée. Its missions are:

- to promote the use of diffraction and neutron spectroscopy,
- to welcome and assist experimentations,
- to develop some research on its own scientific programmes.

Classified as a " Large Installation ", LLB is part of the European NMI3 program (The Integrated Infrastructure Initiative for Neutron Scattering and Muon Spectroscopy), granted by the European Union.

Every year, 400 experiments are performed at the LLB, 70% by French teams and 25 % from European ones.

The LLB has developed a general system for data collection and storage called Tokuma, unlimited in time easily accessible on request. The traditional data format at the LLB is XML but for the instruments generating high amount of data, Nexus format has been chosen.

The LLB support software for data treatment analysis for all type of experiments since many years, which can be download either on the LLB website or on request.

Dr. Stéphane Longeville is in the Biologie et Systèmes désordonnés group in the Laboratoire Léon Brillouin of the CEA. The group studies the structural and dynamic properties of protein folding.

2.3 Consortium as a whole

The participating organisations comprise a very substantial part of Europe's Research Infrastructure in a number of strategic research domains including materials science, bio-medical, nanotechnology, energy applications and fundamental sciences. The common infrastructure of standards and policies agreed between these organisations will therefore quickly become established as a model for similar facilities.

The participants are already working together as the PaNdata consortium (http://www.PaNdata.eu) whose aim is to construct and operate a shared data infrastructure for neutron and photon laboratories. This consortium, through its own independent activities and through the EC-funded Support Action PaNdata Europe, is already working effectively in areas of data policy, user access, data analysis software etc., but there is a need and opportunity to go further in establishing other bases for the infrastructure, which are the aims of the present proposal.

The participants provide the necessary skills, variety of experience and outreach capability, paired with a strong focus on common objectives, which will enable effective work and rapid progress within the available budget.

The currently available (and potential future) data to be made available from the participating organisations is substantial. This provides the necessary and demanding test beds for standards development and, later, their embodiment in supporting technology and roll-out as services.

The Research Institutes involved in this consortium form concentric rings of participants. Seven of the participants (including all work package leaders) form the core for delivery of the project. This is supported by four institutions with lower levels of involvement who are involved directly in the consortium to deploy, test and evaluate the developments to support the sharing of resources across the community. Knowledge exchange activities will then disseminate this to further institutes within Europe and beyond from this critical mass.

The geographical pairing of some of the neutron and photon facilities provides the required complementarity for enhancing close collaboration across disciplines whilst the larger group of photon and neutron sources provides particularly deep penetration into this community, representing a large part of this community within Europe.

The large and overlapping user bases of the research institutes mean that the benefits of the project are immediately transmitted to many thousands of scientists, covering scientific disciplines from medicine to fundamental physics to aeronautical engineering, and distributed through almost all European countries, thus contributing to better science and new science.

The high international standing and influence of the organisations gives the greatest possibility for the results of this project to set the European, and potentially international, standards in this area.

Many of the key personnel in this proposal are regular users of neutrons and photons in performing their own science. As such, they are well placed to provide a well-informed opinion of what scientists actually want from such facilities, beyond access to instrumentation.

The STFC e-science department adds substantial computing expertise, and is uniquely well placed to understand their particular requirements and mode of working. It is extremely well connected to European e-science activities and can hence provide maximum benefit from these to the project.

The involvement of the core partners is divided across the workpackages depending on their current expertise and in order to concentrate the expertise available and form focussed teams developing the common basis through liaison with the other partners.

The Joint Research Activity work packages will each involve a small number of partners (normally three) forming focussed teams concentrating on the particular theme of the activity, and utilising the expertise of the involved partners. The Service Activities will of course involve all partners, as the aim is to evaluate and roll out the developments across the consortium as far as possible. All partners will likewise be engaged in the dissemination work package.

This proposal is not directly related to industrial and commercial aspects and is not appropriate for the direct involvement of SMEs. In the future there is potential exploitation by companies offering added value services based around the repositories, in the same way that companies currently offer database products and other software services associated with repositories of crystallographic data. Industrial and commercial users of the participating facilities will benefit in the same way as all other users. The main benefit to the EU in a commercial/industrial sense comes from improving the 'time-to-market' for information obtained from these RIs, whether the 'market' be publication in the open scientific literature, patenting of results that can be readily exploited, greater exposure of information (improved dissemination) or enabling improved exploitation through the easy overlay of complementary information.

By improving the 'time-to-market', we enhance Europe's position in the increasingly-competitive world 'scientific market'.

2.4 Resources to be committed

2.4.1 Complementary resources

For each of the participating facilities, the generation of scientific data is their main line of business, thus this project will complement an ongoing and substantial investment in the *production* of the data that forms the basis of the repositories. They will provide all of the underlying necessary IT support for maintenance of the repository and hardware systems both during the project and in the future. The facilities will mobilise the following resources to complement and integrate with the work of the project:

Infrastructure Development. Each facility currently maintains a programme of infrastructure development to support its scientific activity. For example, STFC's e-Science Centre has a team of ten supporting data management for science facilities, providing services to ISIS, Diamond and the Central Laser Facility (CLF). These teams will collaborate with the project to provide software infrastructure and tools which integrate with the common infrastructure.

User Offices. Each facility maintains a user office of dedicated staff with a managed user database, each of some 2000–10000 registered facility users. The user offices register users with the facilities, supply them with appropriate authentication and authorisation, and manage the proposal approval processes. Currently, several facilities use an Oracle database to manage this information. These databases will provide information for the pan-European user identification system. The User Office teams will be the prime users of the resulting integrated AAA service.

Data Acquisition. Each facility has a number of teams supporting beamlines and/or instruments which maintain the data acquisition systems and assist the scientists in the generation of data. The project will work with selected teams at each facility to access and integrate data acquisition systems.

Data Analysis. All partners provide substantial support for the intermediate data analysis and treatment, including high performance computing. For example, STFC provides access to the SCARF computational cluster and the UK National Grid Service to ISIS and DLS. Further, specialist teams provide advice and access to analysis and visualisation software.

Data Management. Each facility operates data storage systems to store and manage data generated from the facilities. These data storage and management capabilities will be made available to the project forming the basis of the metadata catalogues and common data holdings.

The following table gives an indicative estimate of the net cost of existing deployed resources on these activities at some of the participating facilities.

	User Office (k€/year)	Data acquisition (k€/year)	Data management (k€/year)	Data analysis (k€/year)	Infrastructure development (k€/year)
ISIS	220	400	300	400	150
ESRF	340	900	400	630	150
ILL	300	600 (ICS service)	180	300	120
DIAMOND	200	600	160	100	120
PSI	300	1100	300	600	100
DESY	200	600	150	200	300

2.4.2 Aggregated resources

The partners have a substantial existing commitment to the development of the constituent components envisaged by PaNdata–ODI, although this is currently targeted at the specific services and user-base of each facility separately. This project will leverage this investment to provide integrated services across the facilities and for the wider community of users across Europe. This will deliver economy of scale as well as facilitating access to existing users of one facility who may be new to another, and also to potential new users who may otherwise have difficulty accessing the resources of the facilities. Thus more and better science will be enabled across Europe.

The dedicated effort of the PaNdata–ODI project will be directed to activities which will benefit all the partners. In particular, the Service Activities will establish common technology and processes across the collaborating facilities. This collaborative effort will engage with existing teams and the active research communities who are eager to exploit this interoperability. The collaboration will thus foster productive exchange of knowledge and propagation of best practice.

These benefits will apply across the whole photon and neutron facility user community which forms a significant aspect of the European Research Area. This makes these collaborative endeavours appropriate to be financed at a European level.

2.4.3 Contribution from the Partners and European Commission

The PaNdata–ODI project will support the development of new technologies and services and their deployment into the operational environment of the collaborating facilities. The operation of those services will then become incorporated into the normal procedures of the facilities.

The EC is asked to support the additional cost of establishing this common technology across the consortium, whereas the partner's own resources will be dedicated to providing the Services for their own user communities. The EC are asked therefore to support 75% of the cost of the Joint Research Activities but only 50% of the cost of the Service Activities, the remaining costs being covered from the partners' own resources.

A contribution is also requested from the EC for the cost of effective collaboration such as travel and management of the project itself. The sums allocated to support this are sufficient to engender a close collaboration between the teams and to manage this tight-knit and focused project.

3 IMPACT

3.1 Expected impacts listed in the work programme

3.1.1 General aspects

The European dimension. As described earlier, the future ICT related challenges will affect all neutron and photon facilities in a similar way. Hence, the most obvious impact of the proposed project is that, for the first time, these challenges will be addressed in a cooperative way by the participating facilities. This is highly significant as, except for the ESRF and ILL, these facilities are financed and focused nationally; which explains why, up to now, many developments have occurred on a purely national scale and why European funding, through this proposal, is required to fully exploit the opportunity to bring together national and international facilities.

Apart from the immediate benefit from the more efficient use of the facilities and e-infrastructure, this project will significantly reduce the number of parallel developments and ease the adaptation of integrated frameworks. The cooperation has already triggered discussions on synchronization of hardware and software investments that will have a positive impact for the facilities and their users and should lead to a significant reduction of investment costs in the future.

The benefits of the cooperative approach proposed here are obvious. Firstly, as the majority of the European neutron and photon facilities will be participating in this project, it is almost certain that the solutions developed will be adopted by *all* European neutron and photon facilities in due course by pure central attraction. Furthermore, the new Free-Electron Laser facilities will face similar challenges. They will readily profit from the outputs of this project. This will, in turn, have a very strong influence on future developments by similar facilities outside Europe.

This cooperation will also have benefits beyond the immediate scope of the project. For example, although this project focuses on software infrastructure, the many regular discussions between the facility decision makers to prepare this proposal have already led to broader discussions, such as the synchronisation of hardware investment decisions, which are positive for the facilities and their users.

The user dimension. The importance of central facilities to world-class science is obvious, yet many potential users fail to visit and exploit them. Many experimentalists accustomed to working in university laboratories perceive that there is an ‘activation energy’ associated with applying for beamtime, visiting a facility, using facility resources and interacting with a facility post-experiment. All the facilities represented in this proposal have made significant efforts in recent years to disavow potential users of such pre-conceptions, and the service activities outlined here represent a *significant step forward* in lowering the ‘activation energy’ still further. This is critical, as facilities are increasingly targeting, and benefiting from, a changing user base, and in particular from users who use facilities as only one part of their overall research programme. A good example is that of the macro-molecular crystallography user community—often the largest community at photon sources—for whom the experiment at the facility is only one step in the experimental chain. The services targeted in this project will have a significant impact upon the user experience when using a range of central facilities.

New scientific opportunities enabled by ‘virtual laboratories’. In this project we aim to provide an infrastructure, which records, maintains, and extends the relationships between scientific experiments, raw data, derived data, software, people, places, times, results,

publications etc. In this way, we are empowering researchers not only to improve the exploitation of their own scientific data, but also to leverage the knowledge of others at all stages of the scientific process.

In the same way that the connectivity provided by the WWW has resulted in ideas and applications beyond any that could have been predicted at the time when it was introduced, it seems clear that the rich connectivity envisaged within this proposal will catalyse lines of scientific research that we simply cannot predict. We provide here only two simple examples of the way in which the infrastructure might be utilised.

Cross-facility, cross-discipline data searching

Consider a small protein molecule where a user has information on the positions of the non-hydrogen atoms in the crystal structure. The scientist wishes to refine the structure but requires more information for a successful refinement. Searching the facility catalogues, they find that is has also been studied by neutron single-crystal diffraction (yielding information on the hydrogen atom positions) and by circular dichroism (CD, yielding information on the protein secondary structure such as alpha helices, beta sheet). They note that the neutron structure factors are available for download and also that the CD work has also been published.

By obtaining the reference, they also find that elsewhere, Nuclear Magnetic Resonance (NMR) measurements have been performed, yielding a set of distance constraints. Pulling all the information together, they embark on a full structure refinement using, for example, the CNS program, yielding a much higher quality refinement than if they had used their original X-ray data in isolation. *It is the ease with which the researchers can locate and access other data that transforms their approach to the refinement*.

Contrast this with the current state of the art, exemplified by some recent research on the early stages of polymer crystallisation using polypropylene, polyethylene and polyethylene teraphthalate that encompassed disciplines from Theory, Materials Science, and the two UK central facilities; SRS and ISIS. The research was hampered by a lack of a central repository for data and associated metadata and it was seriously jeopardized as a result. The problems were only resolved when the collaborating researchers found time to meet in person.

A recently published empirical study on data sharing[8] illustrates the problems in a rather drastic way. Solely relying on the willingness of scientists to share data, 9 out of 10 authors essentially refused to grant access to the data, although PLoS' rules on data sharing are actually much more explicit than APA's recommendations. The open data infrastructure proposed by PaNdata with the data policy in place will in contrast provide an automatic ingest of scientific data into data repositories and thereby ensure existence and accessibility of any valuable raw scientific data.

Data overlays

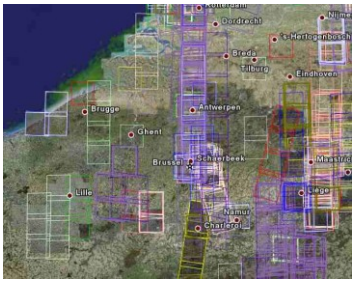
Representing data and results from different scientific disciplines in an easy-to-assimilate fashion should be of great importance to the fundamental understanding of the structure and properties of materials. Moreover it leads to efficient exploitation of the scientific facilities themselves. A vital component is to make the data repositories directly addressable (i.e. using web services the user can achieve programmatic access to data). It opens up the possibility of carrying out very versatile data analysis sessions that touch on a number of data sources. In

[8] http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0007078

the above cross-facility example, diverse data sources were gathered into one location ready for a protein structure refinement.

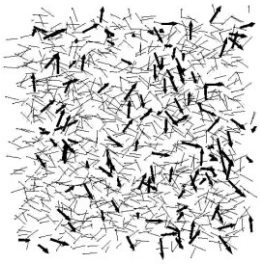
Across disciplines, barriers to communication are reduced through a shared experience of technology and practices. Furthermore, the rapid availability of data from many different types of experimental measurement is crucial to studies of increasingly complex materials and systems. Scientists need to be able to overlay several views of the same objects – a 'Google Earth', at the scale of atoms and molecules. (See figure below.)

GOOGLE EARTH IMAGE OF BELGIUM

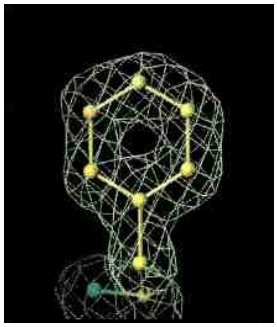


OVERLAID WITH POPULATION CENTRES AND SATELLITE COVERAGE

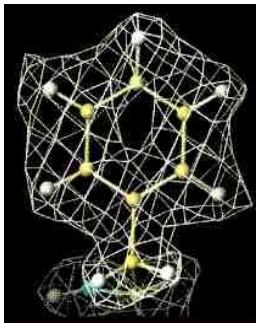
ATOMIC STRUCTURE OF A METALLIC GLASS (AS USED FOR SECURITY STRIPS IN SHOPS) DERIVED FROM TWO SETS OF EXPERIMENTAL DATA FROM A SYNCHROTRON



OVERLAID MAGNETIC STRUCTURE OF THE SAME GLASS DERIVED FROM DATA FROM A NEUTRON SOURCE



STRUCTURAL ELEMENT OF MYOGLOBIN DERIVED FROM SYNCHROTRON DATA



OVERLAID WITH HYDROGEN POSITIONS DERIVED FROM NEUTRON DATA

Integration of systems allowing overlaying of information from different analyses

The atomic scale images shown in the figure are rare examples which can currently take years to achieve. If Europe is to really exploit its large scale multidisciplinary research infrastructures, to significantly improve the 'time to market' of the research results they produce, and to enable new research methodologies, then the implementation of a modern and common data infrastructure is essential.

3.1.2 User catalogues and data catalogues

An integral component of the project is an authentication and authorization system that is normalised to include scientific users across the collaborating facilities and able to interoperate with similar systems across the ERA. The system delivered here is not to replace the local systems of the individual facilities, but rather to allow these systems to interoperate such that individual scientists can be identified on a pan-European level. One major benefit of this is that individuals will be able to seamlessly access all their resources at any of the facilities without having to authenticate themselves against the different systems in place at the participating partners. Other benefits include ease of maintenance which arises from the elimination of multiple entries for particular users and the ability to follow individual scientists as their careers progress through different roles, at different facilities, and across national boundaries. It therefore removes one significant obstacle to the coordination of research policy and practice across Europe.

The implementation of a reliable pan-European photon and neutron user catalogue and portal will for the first time offer corporate-ID functionalities. It will allow exciting new possibilities, such as users being made aware of research opportunities, or allowing for largely simplified conference organisations, etc. A very important aspect of federated user authentication and authorization in the context of distributed data access is that many existing solutions for user authentication can be adopted for the specific needs of the photon and neutron community. User catalogues play a critical role in overall data management schemes. If controlled access to files and resources (e.g. CPU) is to be provided in a coherent and logical fashion, it is essential to verify the identity of the person accessing those files and resources. This is particularly true when using the 'single sign on' approach as envisaged in this proposal.

Further IT-based ways of experimenting are on the horizon and increasingly discussed by the scientists. One example is remote experiment access, where e.g. senior scientists are able to participate online in a running experiment and coach younger scientists on site. Another case is Fedex-type experiments, where the duration of some experiments are getting so short that it is more efficient to send the samples to the facility and the measurement be performed by a local scientist. These experimenting techniques again require unique user identification.

The overall effect will be to promote and ease mobility of users throughout the facilities, resulting in better use of the facilities (and facility resources) and promoting collaborations across sites. It will provide a significant component of a wider European researcher authentication and authorisation system.

It is not realistic to replace within this project the existing local user databases by a single central European user data base, especially in view of the many local tools developed at the various facilities, e.g. automatic access to experimental hutches for users from currently running experiments. Instead, a combination of a centralised and federated approach is planned, where only a subset of the personal coordinates is shared between the facilities.

Apart from users, the other service aspect of the project concerns data. Often described as metadata databases (i.e. databases that keep track of pieces of data that describe other data) the deployed data catalogues will capture details of data files generated by facility instruments during experiments, during subsequent analysis, and through to publications. At their most basic level, they provide a quick and convenient way for users to search for and retrieve their own experiment data. However, such access is merely the tip of the iceberg in terms of the potential benefits for the scientific community at large which will accrue from the participating facilities adopting common data catalogues. Some of these are outlined below.

At the time of the proposal submission, users can search across facilities to see if their experiment or related experiments have already been performed or if the data they are seeking is in fact already publicly available. This is very helpful for the proposers in writing the state of the art section of the proposal. Members of a beamtime review committee can perform similar checks to put the proposed experiment into perspective e.g. is a proposed experiment effectively a duplicate of a previous experiment, or a direct competitor of a similar experiment proposed by a different group?

During the experiment, data produced by an instrument will become instantaneously accessible to authorised members of the experimental team, regardless of their location in the world, enhancing the prospects for immediate analysis and assessment of the data. This in turn leads to a better steering of the experiment. Data produced at the experiment will be 'annotated' with valuable metadata, greatly enhancing its long-term value for owners and those who wish to access it once it becomes publicly available.

Post-experiment, users will be able to access their data easily from their home institutions via a web (services) interface. They will be able to associate other data (e.g. reduced or derived data) with their own raw experimental data by using the data catalogue. In most cases, it is this reduced data that is most useful in the data analysis stage, and thus the ability to associate it with the original experimental data for subsequent search and retrieve by the users (and others) is a significant advance.

Taken together, the above benefits point towards a major change in the way in which users will interact with their data before, during and after a facility experiment. Collaboration between users in a group will be eased via shared access to files and information, especially when it is delivered in near real-time. This can only improve the way in which experiments and post-experiment analyses are performed, leading to the delivery of results in a more efficient and timely manner with potentially better quality.

3.1.3 Provenance and preservation

Data catalogues as outlined above provide a valuable capability to locate particular data sets seen as a 'snapshot' of the scientific process of which they are one of the outputs. The project aims to go a step further and, by representing the *provenance* of the data, link it into the context of the process—the lifecycle of scientific endeavour. This will have a number of impacts. It will become possible to validate published results, linking back through the analyses to the original raw data gathered from the instruments. This is not just a safeguard against errors or fraud, but enables the application of improved analysis techniques as they become available, without needing to repeat the entire experiment, by securely establishing dependencies and derivations from preceding data sets in the chain. Thus efficiency will be improved along with the reliability of the results of such subsequent analyses.

To illustrate the importance of the provenance with an example: scientists investigating the structure of a protein in complex with a ligand encountered that the published small molecule structure of the ligand was an mirror image of the structure deposited in the Cambridge Structural Database (CSD). If the data had been properly curated, the issue could have been resolved within minutes. Lacking the data leading to structure deposition, the scientists had to redo the small molecule structure delaying the publication of the liganded protein structure by several months, which could have been disastrous in such a competitive field.

The analysis process itself will be made more efficient by automatically annotating the activities. The proposed project will enable tools to be built to support scientific workflows, though the building of those tools is not within the scope of the project. Rather the framework will be established to allow tools to create, read and reason about annotations, which unlocks a great future potential for automating parts of the science process. The aim is to develop an extensible framework for later adoption by other disciplines; the critical mass established in this project (representing 30,000 scientists as users) will ensure a wide impact.

The project's activity supporting preservation will have positive impacts in a number of areas. Some of these are analogous to those for provenance, but with a time dimension. Relationships between data sets will be maintained over time, allowing future work to build on past data, safe in the knowledge of its provenance and reliability. It will be possible to avoid repetition of experiments/measurements, or to understand the degree to which new experiments surpass old ones if based on newer equipment. There will be confidence in old data, and confidence also in the conclusions based on it, as the supplementary information associated with data sets will have been captured during the acquisition/analysis processes.

Finally it is worth noting that a further impact enabled by this work will be the establishing of data as of equal weight to publications, reflecting the effort that goes into its acquisition and validation. Thanks to the tracing of provenance and the representation of preservation information, data will no longer be restricted to a mere stepping stone on the way to the final peer-reviewed publication in a journal or conference; it will have the potential to become a recognised part of the scientific output in its own right.

3.1.4 Scalability

The development of the parallel Nexus API, investigation of parallel file systems and the demonstration on specific use cases has a number of potential impacts.

Impact on scientific research. Currently, essentially all applications in photon and neutron science rely on a strictly sequential data access model. Such a model poses a significant bottleneck for real time analysis and hinders efficient use of advanced computing technology. For example, it is not a major problem to dump a large number of individual files onto disk and do analysis almost simultaneously. However, from a data management point of view this is a poor solution. A logical combination of digital objects, a dataset, should be compiled into a single, self-descriptive, structured data file like a Nexus-file, which however prevents simultaneous analysis.

Development and implementation of a pHDF5 capable Nexus API (pNexus) will overcome such limitations, which will not only accelerate the analysis workflow, but offers new scientific opportunities and lays the foundation for efficient analysis of extreme data rates from highly advances light source like x-ray free electron lasers.

pNexus/pHDF5 will further allow to couple directly multiple datastreams passed through multi-core architectures to the application or analysis workflow. It can thereby bridge between pre-processing of the raw experimental data (1st or 2nd level trigger) and the data

archival infrastructure, since it allows combining an essentially unlimited number of data streams to be stored into a single file, which includes recording of the accompanying meta-data.

Real time analysis and archival of experimental data are currently somewhat competing approaches. pNexus has the potential to satisfy both sides, since it allows to accumulate data into a format suitable for long term archival and at the same time permits efficient real time processing of the data.

Extending the approach to distributed parallel filesystems, implementing standard protocols like http, will allow researchers to perform integrated analysis independent on the physical location of raw experimental data, at least for some applications.

Impact on standardization. HDF5 has been proposed by the European Commission as the standard format for all binary digital objects. Consequently, it has been recommended by PaNdata and PNI-HDRI partners as their standard data format. However, adaption of the standard data format by developers and user communities is still progressing slowly. Demonstration of the potential of pNexus will accelerate adaption and acceptance of HDF5.

Impact on scalalibity, (cost) efficiency and commercial developments. There are a (small) number of parallel file systems used productively, mostly by the high energy physics (HEP) laboratories, but support and development is far from being optimal. Part of the problem originates from 'commercialization' of the products. HEP communities are therefore investigating alternative solutions. FhGFS and PVFS are proprietary, freely available open source products with a high potential. Investigation of such solutions with respect to pHFD5 capable applications can have a positive impact on further developments, which could also strengthen the European position in this field of technology.

Developments in high performance computing and real time analysis are increasingly utilizing multi-core architectures (e.g. Nvidias Tesla, AMD fusion, Intel tera-scale). These architectures provide very cost- and energy-efficient platforms for compute intensive tasks. The benefit for data intensive applications like photon science experiments is however rather limited, which demotivates rapid implementation of parallelized solutions. Coupling parallel file systems, parallelized data format and applications to multi-core architectures enhances the usability of such platforms for a wide range of scientific approaches, which possibly could help to accelerate both hardware and application development.

The proposed approach is largely independent on hardware architectures and protocols, which permits to easily extend to newly emerging technologies, incorporate data and compute clouds. It will further facilitate scaling up infrastructures to meet scientists' requirements or to cope with upcoming challenges posed by x-ray free electron lasers.

3.2 Dissemination and/or exploitation of project results, and management of intellectual property

The project will develop and implement new technologies for data management at large scale research facilities. The consortium is ideally placed to make effective judgements as to the design and development of these technologies as it includes all major neutron and photon facilities in Europe.

The work package on Dissemination and Engagement (WP2) is specifically directed to Engagement with other initiatives and dissemination of project results, in particular to other research infrastructures.

Dissemination to other research infrastructures will be through contacts and in particular through other relevant I3s, specifically, NMI3 for neutrons which is coordinated by one of the partners, and ELISA for synchrotrons.
There will also be detailed cooperation and information exchange between PaNdata and related ESFRI activities. CRISP, Cluster of Research Infrastructures and Synergies in Physics, is a proposed joint project for INFRA-2011-2.3.4., and will play a particularly important role in PaNdata's engagement with related ESFRI projects if it is funded. CRISP combines efforts from a wide area of different scientific communities, ranging from astronomy through high energy physics to synchrotron radiation. Common to all these different scientific areas is the necessity to deal with raw experimental data with data rates ranging from terabytes per decade to gigabytes per second. PaNdata will both benefit from and contribute to the wider scope of CRISP. Fortunately, several ESFRI projects (EuroFEL, ESRFup and ILL2020) that are participating in CRISP involve PaNdata partners (DESY, ESRF, ILL and PSI) which will greatly facilitate the level of communication and coordination between the projects and offer opportunities to expand the scope of PaNdata by engaging with ESFRI projects like ESS, European XFEL or FAIR.

Such collaborations will further enhance the interdisciplinarity of PaNdata ODI. PaNdata serves already rather different, complementary types of experimental techniques (neutrons and photons), but many aspects of the PaNdata initiative are hardly restricted to these two techniques. For example ion-experiments (e.g. FAIR) are in many respects very similar. The Helmholtz society has consequently created a program named PNI involving all German large-scale photon, neutron and ion facilities (PNI). The PNI consortium is participating in a project named High Data Rate Initiative (HDRI). PNI-HDRI has rather different aims than PaNdata, nevertheless issues like standardization and best practices are naturally of strong interest as well. PaNdata has already greatly influenced the PNI-HDRI project, data format standards will be identical, PaNdata data policies will be proposed and most likely being adopted. On the other hand, PNI-HDRI started to collaboratively develop abstract instrument and experiment definitions based on Nexus application definitions. The collaboration involves PNI partners as well as PaNdata partners like ESRF, PSI or Soleil, and for example Argonnne National Lab (ANL) resp. the Argonne Photon Source (APS).

Rapidly evolving technologies like trusted clouds provide new ways to implement or advance an open data infrastructure. While CRISP for example, intends to participate in developing new technologies and infrastructure, PaNdata plays more the role of a technology consumer: it's most important for PaNdata that the open data infrastructure serves the scientific user communities in the best possible ways, which requires a tight interaction of users, facilities, application developers and technology or service providers. With its active in involvement in many key projects and the open source development of important elements of the ODI, like Nexus or iCAT, PaNdata will promote sustainable implementation of best practices and standards as well as the efficient use and evolvement of future IC technologies.

In terms of the technology and standards developed for the project, the intention is that these are open source to enable the most rapid exploitation by other infrastructures and users. The project outcome will also be disseminated in form of scientific publications and presentations at conferences or exhibitions under the co-ordination of the WP2 leader. The management of knowledge will be carried out according to the usual practice of the participants, within the framework of the Consortium Agreement, engendering maximum public access to results.

The dissemination and publication of results will meet the contractual requirements in terms of disclosure, and the PMB will check for any IPR issues which may arise.

The participating research infrastructures are already very well connected to European and global research infrastructures like EIROFORUM, NMI3, Elisa, EMI and EGI. Sustainability of the collaborative arrangements engendered by this project will align with the EU harmonisation agenda and will be implemented through these and other channels. Early discussion will be held with these organisations to establish common long-term goals and develop an effective working relationship. Of particular relevance for this project are: The European Strategy Forum on Research Infrastructures (ESFRI), The European Research Consortium for Information and Mathematics (ERCIM), The World Wide Web Consortium (W3C), e-Infrastructures Reflection Group (e-IRG), and the EIROFORUM.

3.3 Contribution to socio-economic impacts

The impacts described above in section 3.1.2 should not be underestimated in terms of their potential socio-economic contribution. They are not only about making users' lives easier; the consequences for enhancing the productivity of 30,000 scientists each year are very considerable, both from the points of view of science and human capital.

Increasingly, scientists are using more than one facility to pursue a single scientific investigation. This is primarily to exploit the complementarities of distinct facilities, radiations and instruments, though it is sometimes done pragmatically to increase the chances of be able to carry out an experiment in an era of significant oversubscription of facilities. Experiments performed at different facilities with different environments increase the total experimental 'overhead'—the synchronised approach of the proposed project will provide an enormous step forward in terms of streamlining such ventures.

The new developments envisaged within this project are primarily software investments for the benefit for facility users. The number of users will increase further with the new facilities under construction and those just coming into operation. This user community has the characteristic that the scientific fields are extremely diverse, ranging from classical physics to nanoscience, chemistry, geology, environmental science, life science, structural biology, medical imaging, or even cultural heritage investigations. This means that the know-how and the solutions developed will be disseminated to, and utilised by, many scientific disciplines.

All infrastructures require their users to register in local user databases which form the basis for all aspects of the experiment organisation from proposal submission through to experiment and publication. As mentioned before, users are increasingly performing experiments at more than one facility. Furthermore, postdoctoral researchers, who execute a great many experiments, change their affiliation every few years and the only practical way of keeping track of the many registration changes is to motivate the users to keep registration entries up to date by themselves. Removing the necessity for users to enter registration information separately at each facility impacts positively on both users and the facilities; users benefit from not having to input the same data at multiple sites whilst facilities benefit by being better able to keep track of users. The latter in particular is significant, as small variations in the way in which someone registers may sometimes lead to multiple entries for the same person with significant administrative consequences and the system has to be able to cope with these issues.

The value for facilities and science-political bodies is also significant, both in terms of the way in which facility-generated data can be kept track of, and the way in which a data catalogue system can sit at the heart of various data-driven enterprises, such as accounting,

analysis, archiving and curation. On a European scale, it should be apparent that common data catalogues that can be searched (with appropriate permissions) via a single interface can deliver data that can be used synergistically by end users. A user searching, for instance, for neutron diffraction and X-ray diffraction data from a particular material may find that data and carry it forward into a combined X-ray/neutron analysis. By facilitating this type of data search, which is currently not possible across facilities, we open up a new frontier in data exploitation, with potential economic consequences as the 'time to market' of the results of such analyses is reduced.

4 ETHICAL ISSUES

	YES	PAGE
Informed Consent		
▪ Does the proposal involve children?		
▪ Does the proposal involve patients or persons not able to give consent?		
▪ Does the proposal involve adult healthy volunteers?		
▪ Does the proposal involve Human Genetic Material?		
▪ Does the proposal involve Human biological samples?		
▪ Does the proposal involve Human data collection?		
Research on Human embryo/foetus		
▪ Does the proposal involve Human Embryos?		
▪ Does the proposal involve Human Foetal Tissue/Cells?		
▪ Does the proposal involve Human Embryonic Stem Cells?		
Privacy		
▪ Does the proposal involve processing of genetic information or personal data (eg. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)		
▪ Does the proposal involve tracking the location or observation of people?		
Research on Animals		
▪ Does the proposal involve research on animals?		
▪ Are those animals transgenic small laboratory animals?		
▪ Are those animals transgenic farm animals?		
▪ Are those animals cloning farm animals?		
▪ Are those animals non-human primates?		
Research Involving Developing Countries		
▪ Use of local resources (genetic, animal, plant etc)		
▪ Benefit to local community (capacity building ie access to healthcare, education etc)		
Dual Use		
▪ Research having direct military application		
▪ Research having the potential for terrorist abuse		
ICT Implants		
▪ Does the proposal involve clinical trials of ICT implants?		

I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL