Project No. FP7 – 212348

NFFA

Nanoscience Foundries and Fine Analysis

D4.9

Scheme for Data / metadata repository

Work Package	**No.4**
Work Package Title	**Design of management structure and format of user access for NFFA-RI Centres. Design of NFFA Data Repository and access criteria. Intellectual property issues.**
Activity Type	**RTD**
Lead Beneficiary	**No.1 CNR-IOM**
Estimated P/Ms	**7**
Nature	**Report**
Dissemination level	**Public**
Delivery Date	**Contractual M 30 Actual 31/01/2011**
Task Leader	**S. Cozzini (CNR-IOM)**
Major Contributors	**R. Gotter (CNR-IOM)**
Other Contributors	**R. Pugliese (Sincrotrone Trieste S.C.p.A.), D. Krizmancic, R. Ciancio, G. Rossi, C. Africh, N. Mahne, M. Kumar (CNR-IOM)**

Delivery Slip

	Partner/Activity	Date	Signature
From	CNR-IOM	18/12/2010	S.Cozzini, R.Gotter
Reviewed by	AC&SP Panelists	18/01/2011	All
Approved by	Coordination Board	31/01/2011	All

Document Log

Issue	Date	Comment	Author
0-0	01/12/2010	first draft using standard NFFA format	S. Cozzini

Document Change Record

Issue	Item	Reason Change

TABLE OF CONTENTS

Deliverable D4.9: Scheme for Data / metadata repository

1. INTRODUCTION

1.1. Purpose of the document

The purpose of this document is to the design the Data Repository prototype where to store scientific data, in the domain of all the data produced in the NFFA centres. Such Data Repository should be compatible at the same time with the maximum usefulness for the referring scientific community, not only in terms of content but also in terms of form (access, readability).

1.2. Application Area

The targets of this document are the members of the NFFA Project, the EC Project Officers, and the general public.

1.3. References

Description of Work (DoW). See at web site:
http://www.nffa.eu/UserFiles/file/Annex_I_DoW.pdf

1.3.1. Objective of Work Package 4

To define the mission and general structure of the future NFFA-RI, including general management of the central RI and of the local facilities, access criteria via quick international review of projects.

Develop schemes for implementing an NFFA-RI data and protocols repository and to make it available to general users. Develop schemes for remote use of NFFA-RI.

Set quality standards of production. Define efficient user access.

1.3.2. Description of work broken down into tasks

The following task was defined in WP4:

T4.9) Design of NFFA-RI Data Repository. Analysis of current **repositories for scientific data and metadata**, for possible remote data analysis codes, for remote use of NFFA-RI and integration in advanced training at universities or other science institutions. Definition of standards of data and metadata (format, remote access via WEB, remote data analysis, preservation and maintenance of data). Technical aspects and realization of a prototype of NFFA repository. Criteria of interoperability based on open standards.

2. EXECUTIVE SUMMARY

Data (the content considered to be the thing lodged in the repository and of primary interest for deposit and use) and **metadata** (all the descriptive part which contains other information pertaining to that content); **users** (generating data) and "**metausers**" (accessing already stored data) constitute the chain above which to design the NFFA Data Repository (DR).

Needs and objectives, which have been discussed more by a "user driven" than a "technology driven" approach in a Coordination Board Meeting in 2009 in Barcelona, require that the NFFA-DR should be:

- less invasive as possible for users (automatic procedure by instruments)
- useful for external references (also for industries and technological districts)
- well integrated in the NFFA management (e.g. for the Technical Liaison duties)
- useful for the implementation of the common platform of metrology and protocols standards
- easy to connect to other e-infrastructures

On this basis the most important technical guidelines have been addressed in order to characterise the few main aspects of the NFFA-DR:

- The **data** reside in **file archives**. The **metadata** are managed using a **database**.
- Metadata management and search criteria will be performed by both **keywords** and **semantic search**.
- Extensive support for metadata management and data processing.
- A flexible architecture that can be adapted to multiple scenarios (transparently extendible to other e-infrastructure).

The NFFA database is therefore a relational database where basic tables are organized in three levels.

1. parameters characterising the object of the research which will be useful for metausers external access:
    - Sample
    - Experimental technique
2. search criterion characterizing the user access.
    - Proposal ID
    - Description
    - Users
3. experiment (single measure, synthesis, fabrication or simulation):
    - Data address
    - Data description
    - Metadata (made of keywords and free text which will be handled with semantic search)
    - Useful data flag (singling out final useful data and not preliminary ones carried out in order to optimise the experiment)

A fundamental role is played by the **keywords management**. As much as possible all the input parameters for these tables should be keywords that can be easily recognised in the search operation and by which a smart priority displaying order can be generated. Special keywords will be used also in the metadata description in order to make it as standardized as possible. For these reasons the keywords sets cannot be arbitrarily free but they must be managed in a proper way, by the Technical Liaison, taking into account for on demand requests by the users as well as for the need to keep it as much light as possible for achieving a simple and quick compiling of the metadata input forms.

In order to effectively face the challenge of the first Data Repository for protocols and standards in nanoscience and nanotechnology, the development of a **small size prototype** has been decided. A Data Repository prototype where to store scientific data, in the domain of all the data produced in the NFFA centres. An analysis of current repositories for scientific data and metadata, for possible remote data analysis codes, for remote use of NFFA-RI and integration in advanced training at universities or other science institutions, has been carried out. Many projects do exist which produced a sizable amount of technology for e-infrastructures and scientific data repositories. We analysed several of them and in some depth Archer and D4science initiatives and finally ICAT. After such an analysis we decided to leverage on the **ICAT** system to build and model our DR prototype for NFFA project as a proof of concepts.

A demonstrator on the top of existing e-infrastructure located at Elettra Synchrotron was then designed and build. The database was organized by means of I-CAT technology. The Demonstrator was connected to three different kinds of experimental facilities to validate the system on three real user cases scenarios:

- a SEM instrument
- an APE synchrotron beamline
- a quantum/espresso simulator to mimic a theoretical facility.

The overall system is available on line at https://nffa.grid.elettra.trieste.it for demo purpose.

The prototype (demonstrator) implements existing technology and standard available.

The demonstrator was made available to some early adopters in order to assess the overall performance of the tool. The first impression is that the tool does fulfil some of the need expressed in the right way, while some others are still far to be fulfilled.

One major point we noticed is that, despite the fact that ICAT tools and proposed standard is widely diffused and available still the product has some constraints (for instance the standard is not well suited to implement the keyword management as above specified and no semantic research on data are allowed at the moment).

The design of the NFFA-DR and the deployment of the demonstrator witness that available tools are still a little bit immature to accomplish all the requirements coming out from the NFFA design study.
Some effort are therefore required implement the improvements and enhancements of this demonstrator to make it become a real production Data Repository for NFFA.

3. THE STRATEGY FOR THE NFFA DATA REPOSITORY

3.1. Data and Metadata, Users and Metausers

In collection of items, items themselves are the units of information as a whole that can be than stored and managed in a data repository. We can often distinguish two sorts of information in an item: its content (**data** itself) and metadata. Content is what is considered to be the thing lodged in the repository and of primary interest for deposit and use generally a data file. **Metadata** are descriptive and contain other information pertaining to that content. Metadata can be extensive, and varied – such as descriptive information, annotations, indexing, classifications, technical information about the content's file formats, and more.

On the other hand we can also consider **users**, that is the researchers generating data when using the NFFA facilities and **metausers** accessing the NFFA Data Repository in order to search, view and use already stored data. This second access will constitute the novel paradigm for nanoscience data, particularly when a standardized platform of metrology and protocols is intended to be set up.

3.2. The concept of Data Repository

A broad and general definition of scientific data repository focuses on data storage. However, the means to retrieve and analyze data, to extract, transform and load data, and to manage the data dictionary are also considered essential components of scientific data repositories.

Thus, an expanded definition for scientific data repositories includes tools to extract, transform, load and search data into the repository, and tools to manage and retrieve metadata.

Scientific Data Repositories are designed to facilitate reporting and analysis and not to complicate life for scientific users. It is therefore fundamental in the design process to take into account the user point of view more than the technology aspects.

3.3. Definition of needs and objectives addressed to the NFFA User Communities

NFFA User communities are the ultimate responsible to define correctly what is actually needed within a DR. The main concept adopted here is that the design should be *"user driven"* and not *"technology driven"*. After preliminary internal discussion the following have been addressed as the most important needs that should be taken into account in the DR design:

1. Data Repositories should be **less invasive as possible for users**. Ideally all the parameters characterizing the sample environment and the instruments should be acquired in an automatic procedure and stored in the DR. Manually inserted data should be done in a standard and simple way making available smart and well scheduled forms whenever user is requested to operate.
2. Data Repository should be **useful for external references** (in particular for industries and technological districts). This means that access should be possible in an easy and simple interface

which allows smart queries on data by means of keywords management, first, and semantic search, secondly and of course uniform data format, view sorting and priorities.

3. Data Repository should be a tools well **integrated in the NFFA management**. For this reason DR should be able to provide tools for:
    a. Allowing meta-user access to NFFA data
    b. Facilitating user access to facilities
    c. Managing scheduling and real time user flux throughout the facilities
    d. Allowing Technical Liaison duties
    e. Supporting proposals submission and reviewing procedures
    f. Managing IPR issues
    g. Monitoring facilities and instruments workload
    h. Supporting coordination activities in the scientific management
    i. Managing expert users in supporting activities
4. Data Repository should be useful for the **implementation of the common platform** of metrology and protocols standards, in comparing data and protocols and in calibrating instruments
5. Data Repository should be easy to connect to other e-infrastructures

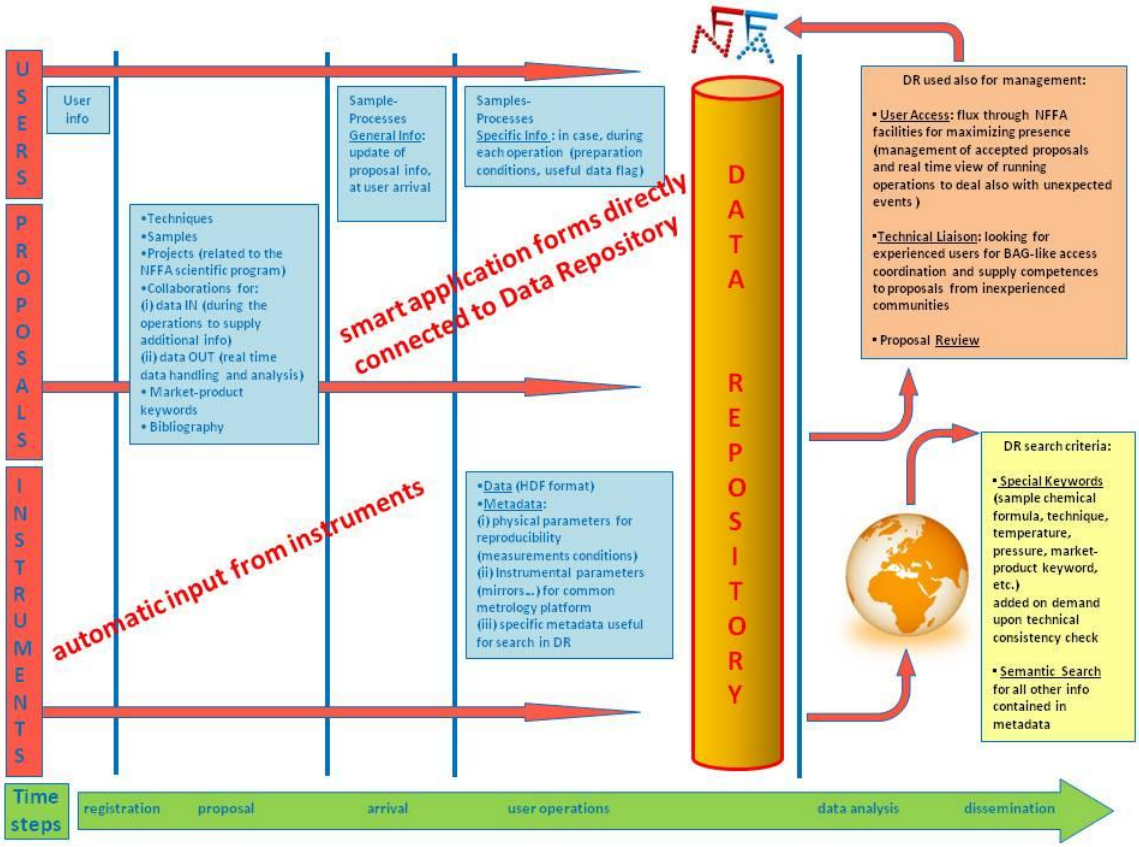

Fig. 1: NFFA Data Repository data flux scheme.

3.4. Guidelines and techniques to be adopted in the NFFA Data Repository

Guidelines and techniques followed in the design of the NFFA DATA Repository are adopted from existing experiences. In Fig.1 a rules-scheme highlighting the input channel (user, proposal and instruments) along with the access time set, is displayed. We can therefore outline the most important guidelines and techniques to be adopted in the NFFA-DR. They comprises:

- Clear separation between data (raw and derived data) and meta data. The **data** resides in **file archives**. The **metadata** is managed using a **database**. There is also a clear separation between

storage schemas for metadata and actual data. This arrangement allows the system to accommodate constant change to the data without affecting the rest of the system.

- Metadata management and search criteria will be performed by both **keywords** and **semantic search**. The former is more efficient but, in order to make it also reliable, a limited well defined set of keywords must be properly managed. The latter is more flexible for accepting new definitions, so in principle more suitable for a continuously evolutionary scientific scenario like nanoscience is, but its implementation is more complex (less reliable) and the search action less efficient (more issues will be anyway found by this method with respect to the keyword criteria).
- Extensive **support for metadata management and data processing**. For searching the metadata, users can use either visual tools to graphically render the search space, predefined queries, or their own SQL queries. For working with the data, users can use the standard analysis routines provided. New analysis results thus produced may be uploaded and imported into the system. A comprehensive set of user interfaces to support the spectrum from the casual, non-specialist user (who can access the system through a Web page) to advanced users (who can create their own tools to access DR).
- A flexible architecture that can be **adapted to multiple scenarios**. We should design a DR where the entire architecture can be transparently extended to e-infrastructure systems with multiple Web servers, processing servers, and a distributed databases.

3.5. Structure of the Database

The NFFA repository is based on an archive where both raw and interpreted NFFA data are stored. All stored data should be available and handled, via metadata processing, for use within the assessment of different workflows/protocols/programs in the various aspects of NFFA related experiments.

The NFFA database aims at establishing a common metadata format and management among the user communities of the NFFA distributed infrastructure. A well organized and unified metadata structure and storage is a key to ensure transparency across all core facilities and will allow a unified scheme for data access and exchange within NFFA partners and users. A well-defined structure is also required to enable the validation of workflows within the project. Finally, the design of the database is such as to allow data to be imported from/exported to other scientific repositories in the more general format and in an as much as automated manner.

The NFFA database is therefore a relational database and it is conceptually organized in more levels where basic tables are built with the following criteria.

At a first level the most important level of search criterion should be identified; for instance, those parameters characterising the object of the research which will be useful for metausers external access:

1. Sample:
   defined by a sample table, e.g.
   - chemical formula: Cu3Au (option, substrate chemical formula:)
   - bulk structure: $L1_2$ crystal (option, substrate bulk structure:)
   - surface structure 111 (option, substrate surface structure:)
   - sample description: thin film, cluster, ….
2. Experimental technique:
   defined by a technique table, e.g.
   - Technique Typology: photoemission
   - Experiment Description: spin resolved photoemission

At a second level we have medium importance search criterion, for instance characterizing the user access slot, which will be useful for user operations, like:

1. Project:

defined by a project table, e.g.

- Proposal ID:

- Description: VOLPE - VOLume PhotoEmission
- Users: Johnson, Robinson

Finally at the lowest level we define the table characterizing the data and the associated metadata and which will be related with the tables of the above levels.

1. Experiment (single measure, synthesis, fabrication or simulation):
   defined by a data table, e.g.
   - Data address
   - Data description: three column spectrum (emitted electrons energy[eV], intensity[counts/s], incident flux [photons/s]), acquisition time
   - Metadata (made of keywords as well as of free text which will be handled with semantic search): sample conditions (temperature, pressure, treatment), measurement geometry (incident angle, polarization angle, emission polar angle, emission azimuthal angle, x-y-z probed zone), photon energy, instrumentation parameters, date and time
   - Useful data flag: singling out final useful data and not preliminary ones carried out in order to optimise the experiment

3.5.1. Keywords management

As much as possible all the input parameters for these tables should be keywords that can be easily recognised in the search operation and by which a smart priority displaying order can be generated. Also when generating data, a set of predefined keywords, displayed in a flag menu in the respective fields of the tables, should be advisable for both simplifying the user interface and achieving a data format as much standardised as possible. Finally, special keyword will be used also in the metadata description in order to make it as standardized as possible. For these reasons the keywords sets cannot be arbitrarily free but they must be managed in a proper way, taking into account for on demand requests by the users as well as for the need to keep it as much light as possible for achieving a simple and quick compiling of the metadata input forms.

If on one hand it will be easy to create a well defined set of keywords for Technique Typology and Experiment Description, as they will refer to the experimental capabilities of the instrumentation available at the NFFA centres, on the other hand a set of keywords characterising the sample prosperities and conditions will be more complex and will probably require a path of subsequent up grades, made in the frame of the Technical Liaison duties.

On a general way, a “new keyword request”, received from any user or from any people of the NFFA staff, will be managed in a coordinated inter-centres way by the Technical Liaison which will decide if including a new keyword or keeping the object of the request as any arbitrary free input in free-fields of the tables. The latter choice will imply that the object will be treated in a semantic way, that is in a less efficient way with respect to keyword labelled objects.

3.6. Issues and final remarks

After internal discussion the above ideas and approaches have been presented to the Coordination Board in Barcelona (March 2009) and there discussed. The overall conclusions are the following:

1. Role of Users is of paramount importance and this aspect should be to take into account for, in all the steps during the DR development.
2. Data standardization: adoption of standard constitutes the basic elements for the development of the NFFA DR.
3. In principle it could be useful to achieve an extreme data log booking, up to a tracing of a complete and detailed path, for instance for a particular sample, form the modelling, to the synthesis, to the analysis, up to the integration in a higher level system.

4. It could be smart to consider a sort of correlation between data originated by different techniques/investigations; for instance the structure of the same sample as measured by a microscopy and as deduced by a spectroscopy.
5. Filtering and quality of data: form the whole raw data up to the significant and useful data to be stored in the DR.

The main issue among Users is about the complexity in the management of the DR from the users. In particular the procedures that should be implemented must be as much as possible close to the present way of working and should not require a large adoption effort. It will be our duty, in the NFFA contest, to design and fit the best interfaces, firstly the input one. The output one will also depend upon the choice of the eventual supporting e-infrastructures and an eventual data customization with them.

But the input interfaces (experiment-DR, see critical points) still remain not well defined, as well as, probably, the output interface (DR-community: we would like a data which is as useful as possible and with an easy and an effective access).

4. ANALYSIS OF PREVIUOS EXPERIENCES/TOOLS

In this section we provide a brief analysis of current repositories for scientific data and metadata. Many important projects at European Level and not only do exist which produced a sizable amount of technology for e-infrastructures. We performed a scouting action in order to see what is already available and which experiences can be considered of interest for us. Our approach is far to be systematic but however we could mention here three example of interest:

- Archer project (http://archer.edu.au/)
- D4Science (http://www.d4science.eu),
- I-cat initiative (http://wwwisis2.isis.rl.ac.uk/DataAnalysis/icat//)

4.1. The ARCHER Project

The ARCHER Project was an Australian higher education initiative, which has developed “production-ready” software tools, operating in a secure environment, to assist researchers to:

- collect, capture and retain large data sets from a range of different sources including scientific instruments
- deposit data files and data sets to e-research data repositories
- populate these e-research data repositories with associated metadata
- permit data set annotation and discussion in a collaborative environment, and
- support next-generation methods for research publication, dissemination and access.

The ARCHER Toolset is a suite of third-party and ARCHER-developed components which assist the collection, storage, management and publication of data from scientific instruments. The released version of the ARCHER Toolset is V1.0, released as Open Source GPL3. No further development has been don after the conclusion of the project.

4.2. D4Science

D4Science **(http://www.d4science.eu)** is a production- level infrastructure serving mainly scientific communities, but which is not biased towards any particular discipline and could have a potential for meeting some of the needs that we have identified for building repositories resources. gCube (http://www.gcube-system.org), on which the infrastructure is based, is a distributed, service-based system designed to support the full life-cycle of modern re- search, with particular emphasis on application-level requirements for information and knowledge management. gCube application services offer a full platform for distributed hosting, management and retrieval of data and information, and a framework for extending state-of- the-art and on-demand indexing, selection, extraction, description, annotation and presentation

of content. Each D4Science VRE, generated using gCube, makes available a grid-based repository to store, share and access information, a grid-based computing environment to efficiently run data analysis services and a reporting tool to publish and share information. gCube be easily extended, as it fully complies with web service standards (SOAP, BPEL, WSRF, WS-* and JSR168 Portal and Portlets.

4.3. ICAT

ICAT (http://code.google.com/p/icatproject/)is a database with a well defined API that provides an interface to Large Facility experimental data and will provide a mechanism to link all aspects of the research chain from proposal through to publication.

ICAT is developed as a collaboration between STFC eScience , STFC ISIS Facility , Diamond Light Source and the ILL.

The ICAT is released as an open source code under the New BSD license (http://www.opensource.org/licenses/bsd-license.php). The code includes libraries under various compatible open-source license.

Data Portal and ICAT API Design

ICAT API is the Grid aware software infrastructure that enables applications to exploit the capabilities of the ICAT system, with potential wider applicability to other research groups and institutions. Data Portal is the Grid aware software infrastructure that serves the Data Search and Retrieval (DSR) requirements of the STFC facilities, with wider potential applicability to other research groups and institutions. Internally it makes use of the ICAT API.

The ICAT API provides a set of web services that the Data Portal uses internally to interact with the ICAT schema. In the primary Data Portal instance, the ICAT API web service capabilities are exposed to the user via a web portal where they can login, search for data across facilities and retrieve the data.

From the information above we considered that ICAT the most closest project to our specific interest and therefore we based our design of the NFFA Demonstrator on ICAT tools as discussed in details in the next section.

5. THE NFFA PROTOTYPE : THE DEMONSTRATOR

In this section we describe the prototype designed and built to manage the NFFA data repository. The prototype called “demonstrator” was designed and implemented by Sincrotone Trieste, at the Elettra Synchrotron site, leveraging on their international experience in setting up e-infrastructure for remote scientific instrumentations. Sincrotrone Trieste, over the years, has acquired great experience in scientific computing, Grid computing, Cloud Computing, e-Science and the so called e-Infrastructure through the participation in Regional, National, European and International projects and experience in solving real daily problems of the scientists. It was therefore the right actor to cooperate with the project to build such a tool and analyze carefully the outcome.

The basic ideas in designing the a prototype are the following:

1. Leverage the existing tools and middleware to avoid to re-invent the wheel.
2. Concentrate on concepts, overall workflow and leave details for future developments
3. Single sign on to all the resources
4. Integrated components that can easily scale
5. Follow other standardisation activities (VEDAC, PaNDATA)
6. Metatada and data curation by means of ICAT tools and possibly automatized via the e-Infrastructure approach.

In the developing of the this prototype we focus mainly on searching metadata by means of keyword, leaving aside at the moment the semantic research that requires more in depth analysis after a testing period of the prototype.

Demonstrator basic structure and detailed usage is described in detailed in Annex 1 provided by Sincrotrone Trieste.

There are three elements (users, proposals, instruments) that will interact with DR. DR is in fact used also for user and proposal management while instruments are providing data (content) and associated metadata.

The technical aspects to take into accounts are outlined in the following steps:

1. Row Data+Metadata creation. Metadata constitutes the set of information characterizing unequivocally the genesis and the use of the data itself.
2. Filtering and processing for setting up the Useful Data, which will be afterward stored in the DR
3. Integration of the Useful Data in such a way that it could be us much correlated and linkable us possible toward other data and other e-infrastructure (input interface), respectively.

5. 1 The Demonstrator: the structure

The NFFA demonstrator is a prototype of the NFFA e-Infrastructure. The architecture of the NFFA demonstrator is sketched in the Fig. 2:

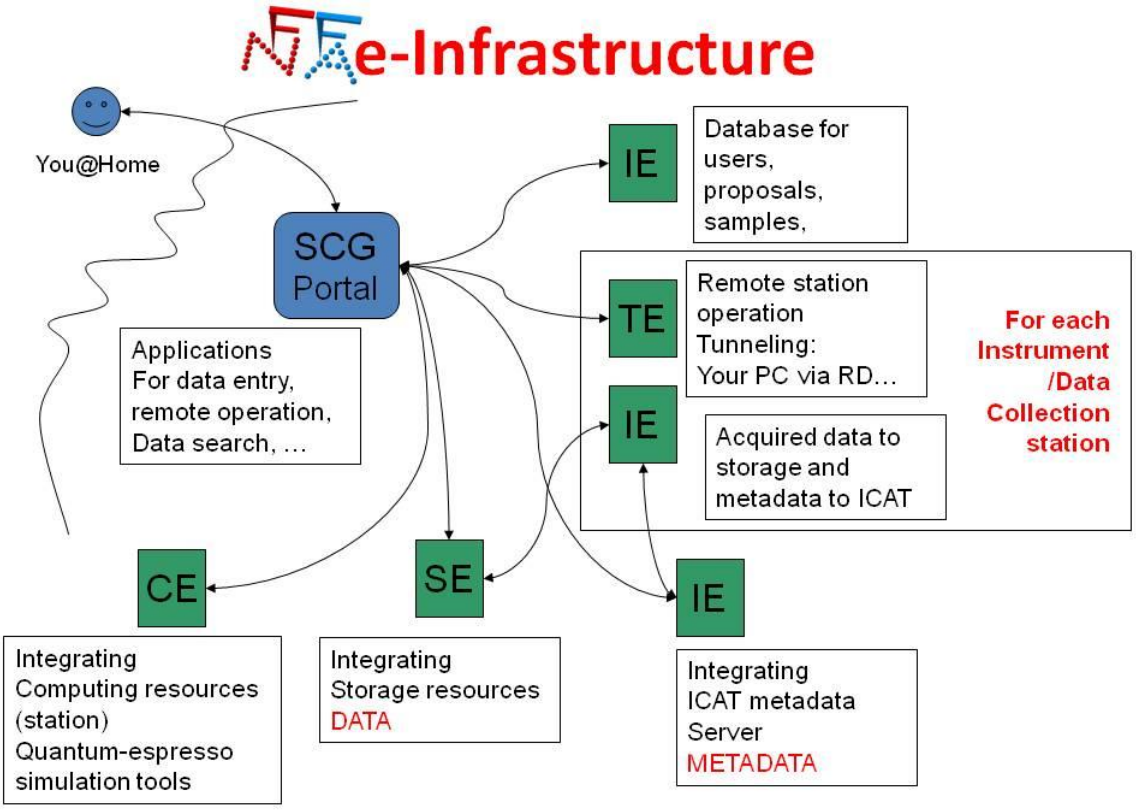

Fig. 2: structure of the NFFA-DR Demonstrator

A generic user of DR will access it through the SCG portal based on the Virtual Control Room product. This is an advance Grid and Cloud portal where several applications are available to the user. These applications are just simple interface to the other core element of the infrastructure, the so-called Instruments Elements (IE) which allow the integration of Instruments and Sensor resources in the traditional computing infrastructure resources like computing facilities (CE) and Storage Systems (SE). At the moment the prototype integrates several kinds of Instrument Elements offering the users the chance to use them through a simple applications embedded within the portal. The portal offers at the moment the following applications (see Fig.3):

a. Data entry applications for proposal submission and handling (submit-proposal and check-my-proposal)
b. An Applications to inject investigation for the integrated instruments (start-investigation)
c. An application integrating the ICAT data portal (open-dataportal)
d. Applications to inject datasets acquired by the integrated instruments (submit-dataset)
e. An application to browse the nffa-storage (browse-nffa-storage)
f. An application which integrates the Quantum Espresso package (quantum-espresso)

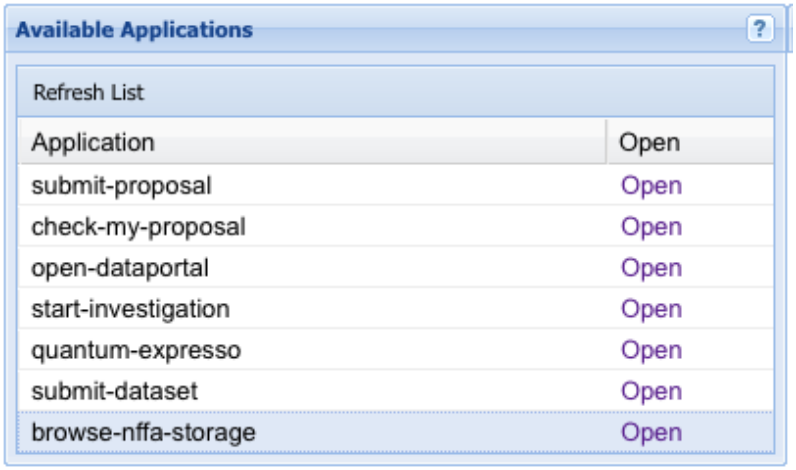

Fig. 3: the pane with the available applications within the portal.

We note that is through the usage of the applications on the portal that metadata are collected (in a semi automatic way) and then stored on icat metadata instrument elements.
Again through the Icat application that integrated ICat data portal these metadata can be sorted out, searched and analyzed.

ICat integration in the prototype required some adaptation of the database design discussed in section 3.4.

The implementation of the prototype has required installation, integration, configurations and developments by the scientific computing team of Sincrotrone Trieste ScpA.
The NFFA prototype and the whole e-Infrastructure is accessible through the VCR at the following address:

https://nffa.grid.elettra.trieste.it/

5. 2 Using the Demonstrator

Users can interact with the data repository by means of the applications mentioned above. Detailed explanation how to use them are available in annex 1 of this tutorial while in the following we will just outline the basic usage:

Step1: a generic user register on the portal and then she submit her scientific proposal to use some experimental facilities within NFFA. She could than check the status of the proposal and give additional information if required to do that. Once the proposal has been approved she could move to step 2.

Step2: User can now start an investigation (start-investigation application) filling in some mask to better define it. Once this step has been completed data acquisition can start.

Step3: User can now use some instrument (like for instance the SEM one) and perform her specific measurements and data acquisitions. Once this is done she could upload data on the nffa storage by means of a simple copy procedure. For this browse-nffa-storage could also be used.

Step4: Data uploaded on the nffa storage can be now inserted in the ICAT database by means of the submit dataset application. Such application will require users to insert appropriate descriptions of metadata associated to the data set.

The above four steps complete the procedure to upload data into the Icat element of the demonstrator. This latter can be then accessed by the users herself to check data and perform keyword search on the data. The application to do that is the "open-dataportal". By means of this application Icat repository can be browsed and searched. The search can be performed on the set of keywords associated by the user during the submit-dataset procedure.

The demonstrator includes also the quantum-espresso application where some simulations can be performed by means of the quantum-espresso package. In this case user can simply upload a an input file where simulation is described and such simulation in then automatically submitted and performed on available computational facility.

6. CONCLUSIONS

The demonstrator was made available to some early adopters in order to assess the overall performance of the tool. The first impression is that the tool does fulfil some of the need expressed in the right way, while some others are still far to be fulfilled.

Some positive aspects of the demonstrator have been reported in this assessment procedure:

- Data repository fully integrated in e-infrastructure
- Data repository follows standard (ICAT)
- Data repository widely available (it requires just a browser and a network connection)

One main concern given by early adopters is that Users are requested to manually load all the data and metadata produced by the experiments into the nffa storage and then submit to the Icat database. The theoretical facility example (quantum-espresso application) is actually a one-step procedure, i.e. the output data produced by the simulation are actually automatically stored in the nffa-storage. Still the user needs to submit the dataset to the ICAT database describing it with appropriate keywords. This at first glance seems quite awkward and not very user friendly.

This two-step procedure can be however partially automatized once metadata available directly by the instruments are uploaded on the storage and automatically parsed by a specific plug-in for each instruments. There is therefore a clear indication how to improve in this direction.

Data security and data accessibility by different kind of users was also reported to be a severe limitation. Again also in this case there are tools and methods already available within the data serviced provided by the portal to setup appropriate security procedure on data uploaded by users and selectively allow access to data.

Finally we report here that, as explained by the Demonstrator developer, a major issue in the development effort was the ICAT tool that actually on a real case scenario did not fulfil all the expectation. In particular we identified the following limitation of the tool:

1. lack of flexibility, particularly for metadata format and keyword management
2. poorly documented
3. more a research project software than stable and robust software to be easily deployed
4. small developing community behind and therefore lack of adequate support
5. lack of semantic supports.

NFFA Data repository needs has been identified and a prototype has been built following closely the requirements suggested by user communities. We remark the importance to try to practically implement a data repository. This allows us to evaluate on ground tools and methods and to spot out clearly limitations and any wrong design of the theoretical design.

The analysis performed so far on the prototype require further testing and works to better define and then implement the improvements and enhancements of this demonstrator to make it become production Data Repository for NFFA.