DELIVERABLE REPORT

WP8: JRA3 – Research on e-infrastructure for data and information management

D8.5

Testbed fully deployed, including DASS services

Expected date

M30

NFFA-Europe has received funding from the EU's H2020 framework programme for research and innovation under grant agreement n. 654360

nffa.eu

PROJECT DETAILS

PROJECT ACRONYM	PROJECT TITLE
NFFA-Europe	NANOSCIENCE FOUNDRIES AND FINE ANALYSIS - EUROPE
GRANT AGREEMENT NO:	FUNDING SCHEME
654360	RIA - Research and Innovation action
START DATE	
01/09/2015	

WP DETAILS

WORK PACKAGE ID	WORK PACKAGE TITLE
WP8	JRA3 – Research on e-infrastructure for data and information management

WORK PACKAGE LEADER

Stefano Cozzini (CNR-IOM)

DELIVERABLE DETAILS

DELIVERABLE ID	DELIVERABLE TITLE
D8.5	Testbed fully deployed, including DASS services

DELIVERABLE DESCRIPTION

This deliverable describes the full IDPR testbed, including the DASS services

EXPECTED DATE	ESTIMATED INDICATIVE PERSONMONTHS
M30 28/02/2018	6

AUTHOR(S)

Rossella Aversa (CNR-IOM), Stefano Cozzini (CNR-IOM) Andy Goetz (ESRF), Kumbhar Snehal Pramod (EPFL), Thomas Jejkal (KIT)

PERSON RESPONSIBLE FOR THE DELIVERABLE

Stefano Cozzini (CNR-IOM)

NATURE

P - Prototype

DISSEMINATION LEVEL

☒ P – Public

☐ PP - Restricted to other programme participants & EC: (Specify)

☐ RE - Restricted to a group (Specify)

☐ CO - Confidential, only for members of the consortium

REPORT DETAILS

ACTUAL SUBMISSION DATE	NUMBER OF PAGES
27/02/2018	14(right-click and select “update the field”)

FOR MORE INFO PLEASE CONTACT

Stefano Cozzini (CNR-IOM) Tel. +39-040-3787508 Email: cozzini@iom.cnr.it

Version	Date	Author(s)	Description / Reason for modification	Status
0	08/02/2018	Stefano Cozzini	First draft	Draft
1	16/02/2018	Rossella Aversa	Added SEM sections	Revision
2	16/02/2018	Kumbhar Snehal Pramod	Added EPFL contribution	Revision
3	22/02/2018	Andy Goetz	Added ESRF contribution	Revision
4	25/02/2018	Stefano Cozzini	Global revision	Revision
5	26/02/2018	Rossella Aversa	Added KIT’ contribution/revision	Revision
6	27/02/2018	Stefano Cozzini	Final revision	Final

Sommario

Sommario 4

Executive Summary 5

1. Testbed structure 6

1.1 The IDRP (idrp.nffa.eu) 7

1.2 The NFFA portal (www.nffa.eu) 7

1.3 The Data Archives of the testbed 7

The CNR-IOM KIT-DM facility: kit-dm-iom.nffa.eu 7

The EPFL Materials Cloud facility: www.materialscloud.org 8

Import of Materials Cloud Datasets in the NFFA-IDRP 9

1.4 The Data Analysis Software Services 10

The CNR-IOM SEM-classifier application: sem-classifier.nffa.eu 10

The ESRF application 11

The STM VM for STM software services: stm-services.nffa.eu 12

2. Conclusions & Perspectives 13

References 14

Executive Summary

This deliverable reports a short description of the final NFFA Information and Data Repository Platform (IDRP) testbed developed by KIT [1], in collaboration with CNR-IOM [2] and Promoscience srl [3], and deployed on CNR-IOM OpenStack cloud infrastructure [4]. The testbed is the enhanced version of the original prototype version described in [5], which implements now also some first Data Analysis Software as Services (DASS).

All the components of the prototype have been deployed as independent virtual machines. Each single element, and the interaction with the others, will be briefly described in Section 1.

A user guide describing the basic procedures to interact with the IDRP can again be found at [6]. These web pages will be constantly updated and is intended to become the manual for the IDRP infrastructure usage and the associated DASS services.

1. Testbed structure

This section briefly describes the components of the NFFA testbed, based on the IDRP concept described in the NFFA proposal (Figure 1) and refined in deliverable D8.2.

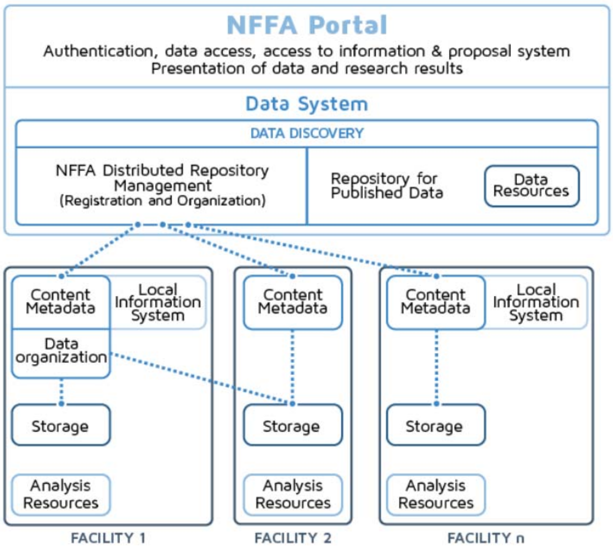


Figure 1: Conceptual architecture, as presented in the NFFA proposal.

The testbed's main services are deployed on the CNR-IOM OpenStack cloud infrastructure. As shown in Figure 2, it is composed of different elements installed on different virtual machines, representing the layers of the architecture. The IDRP is indicated in blue. The orange box represents the NFFA portal, while the grey ones represent external services, e.g. B2SHARE, an external service provided by the EUDAT [7] project to publish measurements, or data management resources located at NFFA facilities. Purple boxes represent instruments producing data, which is stored in local data archives. The data flow is represented by black arrows.

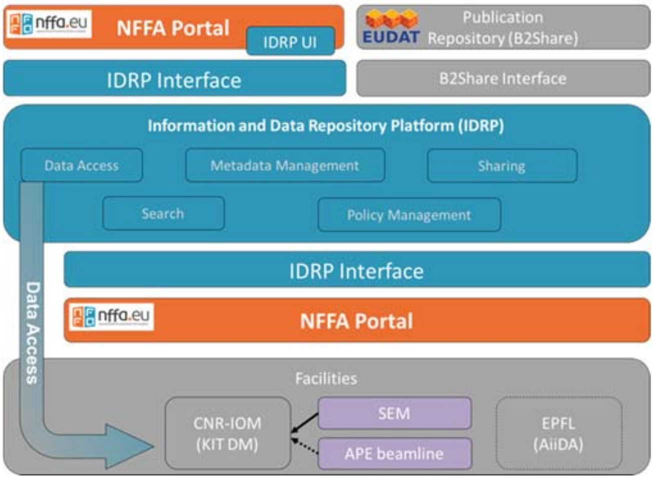


Figure 2: Schematic view illustrating the prototype installation and the interaction among the different components: the IDRP (blue), the NFFA portal (orange), the external service of B2SHARE (grey boxes at the top), the Data Archives at the local facilities (grey boxes at the bottom), and the instruments (purple boxes). All components are installed for testing at the CNR-IOM cloud infrastructure. Dashed lines indicate elements and data flows that have been already planned but not yet implemented or tested.

1.1 The IDRP (idrp.nffa.eu)

The IDRP is the core of the architecture, connected both to the data archive of the local facility and to the NFFA portal [8]. It is now available in its final graphical version at idrp.nffa.eu.

The IDRP exposes several RESTful endpoints providing the following methods:

- to register, manage and retrieve metadata
- to access and link metadata and data stored in the local repositories
- to manage the authorization of metadata and data access
- to interact with the NFFA portal in order to retrieve proposal and user information.

A documentation of all available endpoints can be found at [9].

The main purpose of the RESTful endpoints is to provide a defined interface to the data platform. This interface can be used to e.g. register data assets using custom tools as soon as the data assets are produced or to access data transparently for processing. On the other hand, this interface allows building graphical user interfaces for querying and visualizing data and metadata of NFFA proposals.

In this final installation, a much better user interface has been made available after several discussions with different NFFA users. Such user interface allows scientists to register, modify and retrieve their metadata and data. Furthermore, it allows to share and publish measurements, either to IDRP users or to worldwide using the EUDAT B2Share service. It is moreover fully integrated seamlessly into the official NFFA portal, which serves as single entry point to the information portal as well as to the IDRP.

1.2 The NFFA portal (www.nffa.eu)

The current testbed is now linked to the official NFFA portal, which exposes a REST interface to dialogue with the IDRP to provide authorization and authentication services, and information on the basic metadata related to proposals submitted by NFFA users.

1.3 The Data Archives of the testbed

The NFFA landscape includes many facilities, in which different data management systems are available, e.g. KIT Data Manager [10], ICAT [11], iRods [12], NOMAD [13], and AiiDA [14].

The present testbed includes in this moment three different facilities:

The CNR-IOM KIT-DM facility: kit-dm-iom.nffa.eu

At the CNR-IOM, the data management system adopted is based on KIT Data Manager. The KITDM@CNR instance has been installed on a Ubuntu virtual machine and exposed as kit-dm-iom.nffa.eu

A command line interface (namely click) has been deployed and made available on a public repository [15]. The click has been downloaded and configured both at the facilities and on personal computers, with different operative systems (Windows XP, Linux, MacOS). Through the click, instruments data at the facility is ingested/downloaded to/from KITDM@CNR, which creates the hierarchical metadata structure needed to store them.

An additional plugin for the APE beamline has been already developed, but is not yet fully implemented in the prototype.

The EPFL Materials Cloud facility: www.materialscloud.org

Materials Cloud [16] is a web platform designed to enable seamless sharing of resources in computational materials science, including educational material, interactive tools and virtual hardware to run simulations, up to publishing results in a FAIR-compliant format [17]. Materials Cloud is powered by AiiDA [14], a Python framework to manage materials science calculations, automatically storing the full provenance of data and calculations.

By sharing scientific data on Materials Cloud, not only the results of calculations but every step along the way is made available and is fully downloadable, both as individual files or as a whole database, so that research results can be seamlessly reused. Moreover, the web interface makes it easy to browse and query for calculations and data, as well as it provides a Jupyter-based interface to run simulations in the cloud.

Materials Cloud also provides an “Archive” section that facilitates storing research data, guarantees long-term data availability and helps to share data with community. It assigns a Digital Object Identifier (DOI) to each version of published results to make it easily and uniquely citable. Materials Cloud provides REST APIs that helps to integrate and publish data with other tools. IDRP is one of the services that is being integrated with Materials Cloud. It uses REST API endpoints to get the metadata and curated information in JavaScript Object Notation (JSON) format. The JSON response contains information about published data like title, description, authors, affiliations, DOI, versions, keywords, submission date, etc. for every “Archive” entry to display it on the IDRP interface along with the link to download published results. A snippet of a sample JSON response for entry materialscloud:2008.0001 is shown below (Fig.3).

```
{
   "title": "SSSP library optimized for accuracy and efficiency",
   "description": "Despite the enormous success and popularity of density functional theory, …",
   "authors":[
            {
              "first_names": "Gianluca",
                     "affiliations": [1],
                     "last_names": "Prandini"
                   },
                   …
                 ],
   "affiliations": [EPFL, CH-1015 Lausanne, Switzerland", …]
   "submission_date": "26 January 2018",
   "files": […],
   "keywords": [ "SSSP", "pseudopotentials", …],
   "refs": [],
   "is_draft": false,
   "entry_id": "2018.0001",
   "entry_url": "/2018.0001/v1",
   "doi": "10.24435/materialscloud:2018.0001/v1",
   "last_version": "v1",
   "last_version_submission_date": "26 January 2018",
   "license": "CC-BY-4.0"
}
```

Fig. 3: A snippet of a sample JSON response for entry materialscloud:2008.0001

The Materials Cloud is being developed using modern web technologies: on the server side, the AiiDA API as well as many additional services (online tools, video streaming of lectures and tutorials, the backend of the "Archive" section) are implemented in Python and are exposed via REST interfaces, while the web client uses libraries like AngularJS, D3.js, Bootstrap, JSmol, and JQuery. Materials Cloud is automatically deployed via Ansible scripts in the OpenStack installation provided by CSCS (Switzerland). CSCS is also developing a stack of federated services for authentication and authorization (KeyStone), object storage (Swift) and web services (OpenStack). These services will be extended to CINECA (Italy) and Jülich (Germany) to build a decentralized cloud.

Import of Materials Cloud Datasets in the NFFA-IDRP

Importing Materials Cloud datasets into the IDRP is realized in a pull-based fashion by providing an import dialog at the IDRP Web UI. There are a couple of reasons for choosing this approach compared to a push-based approach implemented by the Materials Cloud platform. In the first place, this integrated approach promises a high level of usability as the user does not have to learn different systems or tools. From the technical perspective, realizing an import by the IDRP improves the maintainability and reduces the effort for developers of the Materials Cloud platform to a minimum.

In order to realize the import, the first task was to create crosswalks between the Materials Cloud metadata model and the NFFA metadata model. The table below shows all crosswalks that have been available for the first version of the importer.

Tab.1: Crosswalk from MaterialsCloud entry to NFFA model

MaterialsCloud Entry	NFFA Metadata Model
Title	Measurement.measurementName
Description	Measurement.measurementDescription
Submission_Date	DataAsset.dateOfCollection
License	DataAsset.intellectualPropertyRights
File.name	DataAsset.dataAssetName
File.size_bytes	DataAsset.dataSize
File.md5	DataAsset.dataChecksum

One can see, that a single Materials Cloud entry maps to an NFFA measurement receiving title and description from the entry. All metadata entities on a higher abstraction level, e.g. proposal and experiment, are either created automatically (proposal) or have to be created and selected by the user who is in charge of importing the Materials Cloud entry (experiment). Typically, only a principal investigator of a proposal is allowed perform imports, thus eligible users have all required permissions to create missing metadata entities.

After creating the according measurement entry, the Materials Cloud metadata is used to extract data asset information. Each Materials Cloud entry contains one or more associated files. During the import the user can select if each file is imported as individual data asset or if the entire set of files should be interpreted as single data asset. In the latter case, only the summed size of all files and the license are used in the imported data asset, whereas individual filenames and checksums are omitted. If each file should be interpreted as single data asset, all these elements are assigned to the imported NFFA data asset.

By realizing the Materials Cloud importer, the user can now register datasets published at the Materials Cloud platform within the NFFA IDRP. This allows the user to link results produced using the AiiDA infrastructure to measurement data captured in the context of NFFA. As the actual data access is still handled via Materials Cloud in a public way, the user can be provided with fully transparent access to remotely hosted data with the benefit to use IDRP's enhanced search features.

1.4 The Data Analysis Software Services

The NFFA testbed includes some examples of Data Analysis Software Services, each one at a different level of maturity and with different approaches not yet fully seamless integrated. In this section we present the first examples.

The CNR-IOM SEM-classifier application: sem-classifier.nffa.eu

The idea of developing a Scanning Electron Microscopy (SEM) classifier originated from the need of SEM users to automatically tag the images they produced in a uniform way. Indeed, scientific data stored into a repository should be FAIR (Findable, Accessible, Interoperable, Retrievable) [17].

Some metadata coming from the instrument (Beamtime, Gun Vacuum, EHT, and so on) have already been present in the images. However, any information about the target material (i.e., what is the subject of the image) were missing.

In collaboration with SEM users, ten categories were identified (Biological, Fibres, MEMS devices and electrodes, Particles, Porous Sponge, Tips, Films and Coated Surfaces, Nanowires, Patterned surfaces, Powder). Different neural network architectures were trained and tested on a dataset of human-labelled SEM images.

The best model achieved was used as the engine of the online analysis service we developed to automatically classify newly incoming images. The user can access the website [18], log in, and upload either a single SEM image or a folder of them. The inference is performed at runtime (a few seconds per image are needed), and the resulting categories, together with the rate (a measure of "certainty"), are suggested. The user can confirm or change the result. A snapshot of the SEM classifier is show in Fig.4.

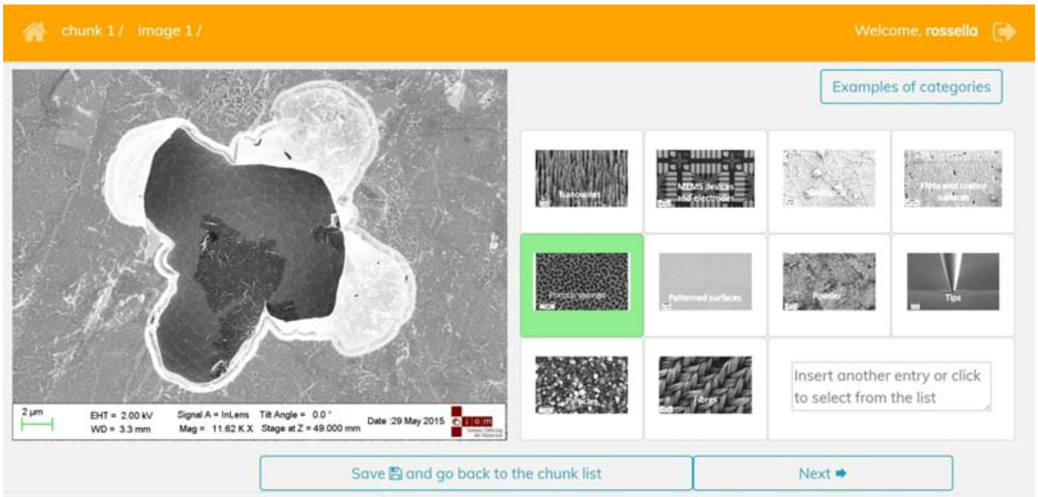


Figure 4: A snapshot of the SEM classifier website. The image on the left has been classified as Porous Sponge (highlighted in green). All the other categories are also shown, together with a representative image.

When the process has finished, the chosen category is included as additional metadata to the image. This has the final advantage that once the images are linked to the IDRP, they can be retrieved by category using the search engine.

The ESRF application

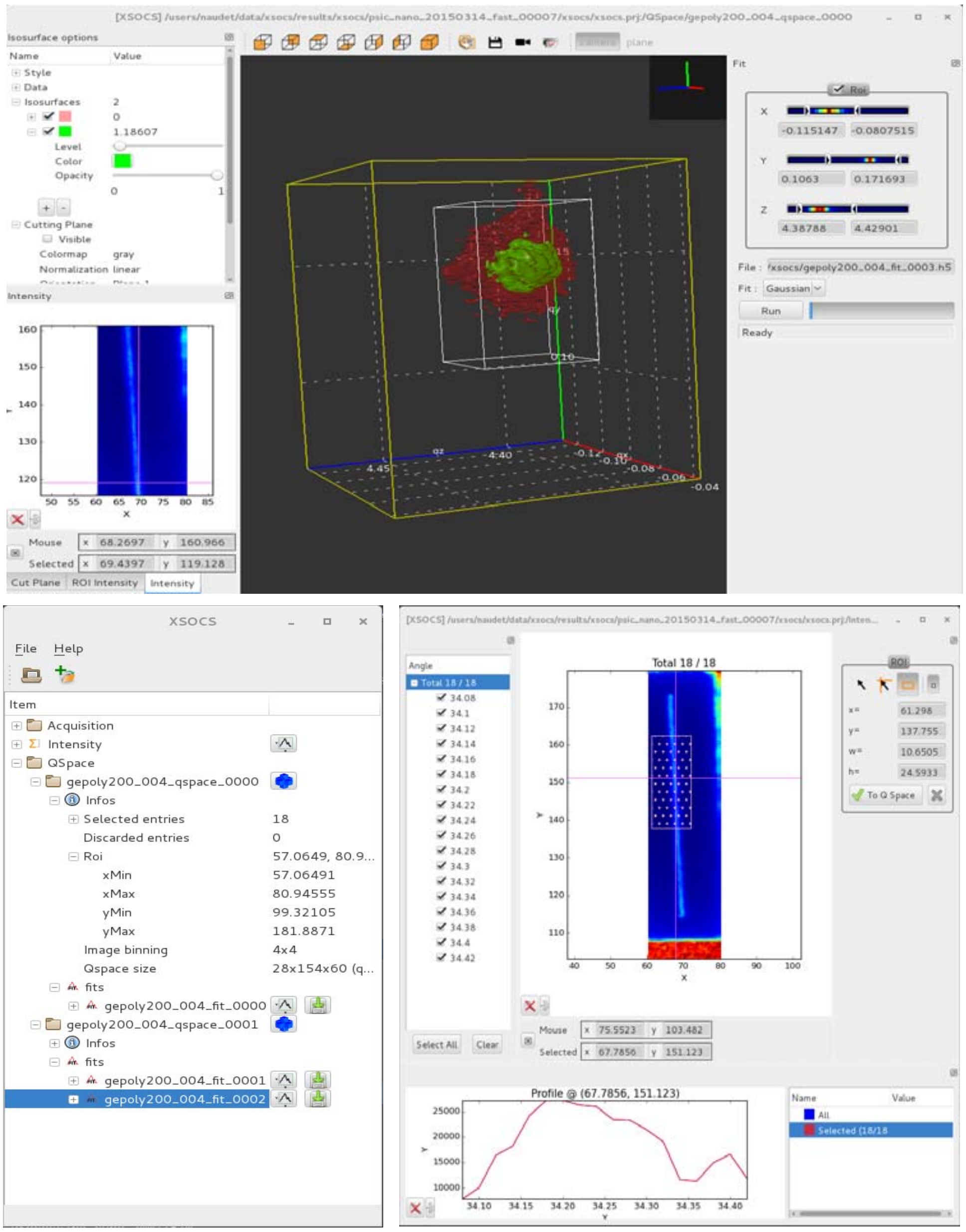
ESRF is a 3rd generation synchrotron providing highly focussed high energy x-rays to up to 40 beamlines for conducting experiments. A large number of these experiments exploit the characteristics of the ESRF source and beamlines to study matter at the nano-scale level. In order to profit from the data collected by the detectors users need good software. NFFA has co-financed the development of a software called X-SOCS for studying strain on the nanoscale level using a technique called KMAP.

X-SOCS aims at retrieving strain and tilt maps of nanostructures, films, surfaces or even embedded structures. It offers the opportunity to get preliminary results directly at the beamline giving the user the opportunity to adapt the planning of the experiments and the measurements with respect to this first set of results. This is of particular importance for the application of such fast scanning methods to *in operando* studies at high temperatures or in gas or liquid environments. The ultimate goal being to offer a tool for showing strain, tilt and or composition map for nano-samples as a remote service for users. Doing so will lead to opening the technique up to new users and applications.

The first version of the software was written by beamline scientists and users. However the software was not very efficient (it took hours to run), was difficult to maintain (because it was built on libraries not supported at the ESRF), and was lacking in advanced 3D visualisation capabilities. The work of the software engineer by the NFFA project was to address these points and make the X-SOCS program maintainable in the long term. The long-term maintenance was addressed by adopting and building on the newly developed silx [19]. Silx offers a library for 1D, 2D, and 3D visualisation routines for reading and writing data in common data formats and it is maintained by the ESRF. Using this library it was possible to have advanced 3D visualisation of reciprocal space in a very short time. The 3D visualisation is more than just a visualization as for a large number of experiments, defining a region of interest on the detected image is key in extracting the correct information. Being able to define such regions of interest after an experiment makes the analysis of the treated data easier and faster. Moreover, the 3D visualization in the reciprocal space gives access to information, hidden with the 2D visualization tools in the initial version of the software. For some experiments, without the 3D visualisation the data would have been thrown away, as it was very complicated to understand the sample and the diffracted intensity distribution. The efficiency of the data manipulation was improved by 2 orders of magnitude by restructuring the code to make it more efficient, adopting HDF5 as data formatting and using a multi-threaded approach. The first step in the data reduction now takes minutes instead of hours. The new program is available on the ESRF cluster and for downloading. The source code is available under an Open Source licence. The program will be easier to maintain in the long term because it is based on supported libraries. Making the program open source means we can hope to attract contributions from other institutes. The X-SOCS software is available from the PAN software catalogue [20].

The success of this project has been strongly appreciated by the ID01 users community and in the words of the beamline scientist it is "*a real game changer for ID01 as a user facility".* This project shows that by investing in software development major improvements can be made in user experience.

The NFFA tested has a specific VM machine where the software has been installed and it is now available to the NFFA user community. Some datasets have also been loaded into the IDRP to provide integration among the data and the analysis service associated.



[XSOCS] /users/naudet/data/xsocs/results/xsocs/psic_nano_20150314_fast_00007/xsocs/xsocs.prj:/QSpace/gepoly200_004_qspace_0000

Isosurface options

Name	Value
⊞ Style	
⊞ Data	
⊟ Isosurfaces	2
⊞ ✓ 🟥	0
⊟ ✓ 🟩	1.18607
Level	
Color	🟩
Opacity	0 … 1
⊟ Cutting Plane	
□ Visible	
Colormap	gray
Normalization	linear

Fit

X: -0.115147 -0.0807515

Y: 0.1063 0.171693

Z: 4.38788 4.42901

File : /xsocs/gepoly200_004_fit_0003.h5

Fit : Gaussian

Ready

Mouse x 68.2697 y 160.966

Selected x 69.4397 y 119.128

XSOCS

Item	
⊞ Acquisition	
⊞ Intensity	
⊟ QSpace	
⊟ gepoly200_004_qspace_0000	
⊟ Infos	
⊞ Selected entries	18
Discarded entries	0
⊟ Roi	57.0649, 80.9...
xMin	57.06491
xMax	80.94555
yMin	99.32105
yMax	181.8871
Image binning	4x4
Qspace size	28x154x60 (q...
⊟ fits	
⊞ gepoly200_004_fit_0000	
⊟ gepoly200_004_qspace_0001	
⊞ Infos	
⊟ fits	
⊞ gepoly200_004_fit_0001	
⊞ gepoly200_004_fit_0002	

[XSOCS] /users/naudet/data/xsocs/results/xsocs/psic_nano_20150314_fast_00007/xsocs/xsocs.prj:/Inten...

Angle

Total 18 / 18
✓ 34.08
✓ 34.1
✓ 34.12
✓ 34.14
✓ 34.16
✓ 34.18
✓ 34.2
✓ 34.22
✓ 34.24
✓ 34.26
✓ 34.28
✓ 34.3
✓ 34.32
✓ 34.34
✓ 34.36
✓ 34.38
✓ 34.4
✓ 34.42

ROI

x= 61.298

y= 137.755

w= 10.6505

h= 24.5933

Mouse x 75.5523 y 103.482

Selected x 67.7856 y 151.123

Name	Value
🟦 All	
🟥 Selected (18/18	

Figure 5: 3D view of the reciprocal space with two isosurfaces (Top), workspace view (Bottom left), acquired intensity view/ROI selection (Bottom right).

The STM VM for STM software services: stm-services.nffa.eu

The CNR/IOM Scanning Tunnelling Microscope (STM) lab is collaborating with us in setting up a dedicated virtual server, which comprises all the analysis software available to STM users. This will provide NFFA-users with an ad-hoc data service for remote analysis on the data acquired in the course of the NFFA visit.

STM researches already developed their own format based on HDF5 files to store images and videos produced by STM experimental techniques available in the lab. In particular, they produce the so-called "fast" file, in a HDF5 format containing metadata and data (1D array to be converted into images using the attributes included in the metadata). From the conversion of the array, the STM video can be recovered. To read and analyse the STM video, a dedicated Python module has been already developed within the lab and these following freely available tools:

- Python module [21]

- other Python modules [22, 23]

- Gwyddion (open-source) through "omicronmatrix" module [24]

The virtual machine has been configured to have all such software on board and it is furthermore connected to easily access the data stored in the IDRP and on the local STM storage resources.

At a later stage also the STM logbook will also be digitalized: the information on the sample preparation can be easily parsed and automatically added as metadata into the HDF5 file before uploading it to the local storage.

2. Conclusions & Perspectives

In this document, we present the enhanced and completed version of the IDRP testbed that includes now the first Data Analysis Software Services

The current IDRP testbed has been developed at KIT, in collaboration with CNR-IOM and Promoscience srl and integrates contribution coming from EPFL, CNR-IOM STM group and ESRF. All the details about the architecture are reported in D8.2 [4], a description of current enhanced functionalities can be found at [6].

Next steps are improving the testbed in terms of number of DASS available and start using the testbed for real production by the majority of the NFFA-EUROPE users. We also plan to work on a more seamless integration of DASS and IDRP.

References

[1] Karlsruhe Institute of Technology website: https://www.kit.edu/english/

[2]: CNR-IOM website: https://www.iom.cnr.it/

[3] Promoscience srl website, http://www.promoscience.com/

[4] CNR Openstack cloud, http://nimbo.escience-lab.org/dashboard/auth/login/

[5] Thomas Jejkal, "Design of the finalized repository architecture", http://intranet.nffa.eu/DocumentRepository

[6] Thomas Jejkal, "NFFA Information and Data Repository Platform", http://ipelsdf1.lsdf.kit.edu/nffa/idrp/manual/index.html

[7] EUDAT website, https://www.eudat.eu/

[8] NFFA portal, http://nffa.eu/

[9] Thomas Jejkal, "Information and Data Repository Platform - RESTful API", http://147.122.7.215:8080/swagger-ui/dist/index.html

[10] Karlsruhe Institute of Technology, KIT Data Manager, http://datamanager.kit.edu/index.php/kit-data-manager

[11] ICAT project website, https://icatproject.org

[12] iRods website, https://irods.org/

[13] NOMAD website, http://repository.nomad-coe.eu/cms/

[14] G. Pizzi et al., Comp. Mat. Sci. 111, 218 (2016) - www.aiida.net

[15] click repository, https://gitlab.com/NFFA-Europe-JRA3/CLI-KITDM

[16] Materials Cloud website, www.materialscloud.org

[17] FAIR principles, M. D. Wilkinson et al., Scientific Data **3**, 160018 (2016), DOI:10.1038/sdata.2016.18

[18] sem classifier, http://sem-classifier.nffa.eu

[19] Silx, https://www.silx.org/

[20] PAN software catalogue: https://software.pan-data.eu/software/132/x-socs

[21] STM access2theMatrix module, https://pypi.python.org/pypi/access2theMatrix/0.2.3

[22] pyMTRX modules, https://pypi.python.org/pypi/pyMTRX/1.9.0

[23] pyOmicron modules, https://github.com/scholi/pyOmicron

[24] Gwyddion module, http://gwyddion.net/download.php#stable-sources