

The NFFA Europe Information and Data Repository Platform

Stefano Cozzini

CNR – IOM



LSDMA symposium, Karlsruhe 29.08.2017



Agenda

- Introduction: NFFA-EUROPE project
- The Information Data Repository Platform (IDRP)
 - Why and What ?
 - The IDRP Prototype
 - One successful use case
- The NFFA-IDRP data policy document
- Conclusions

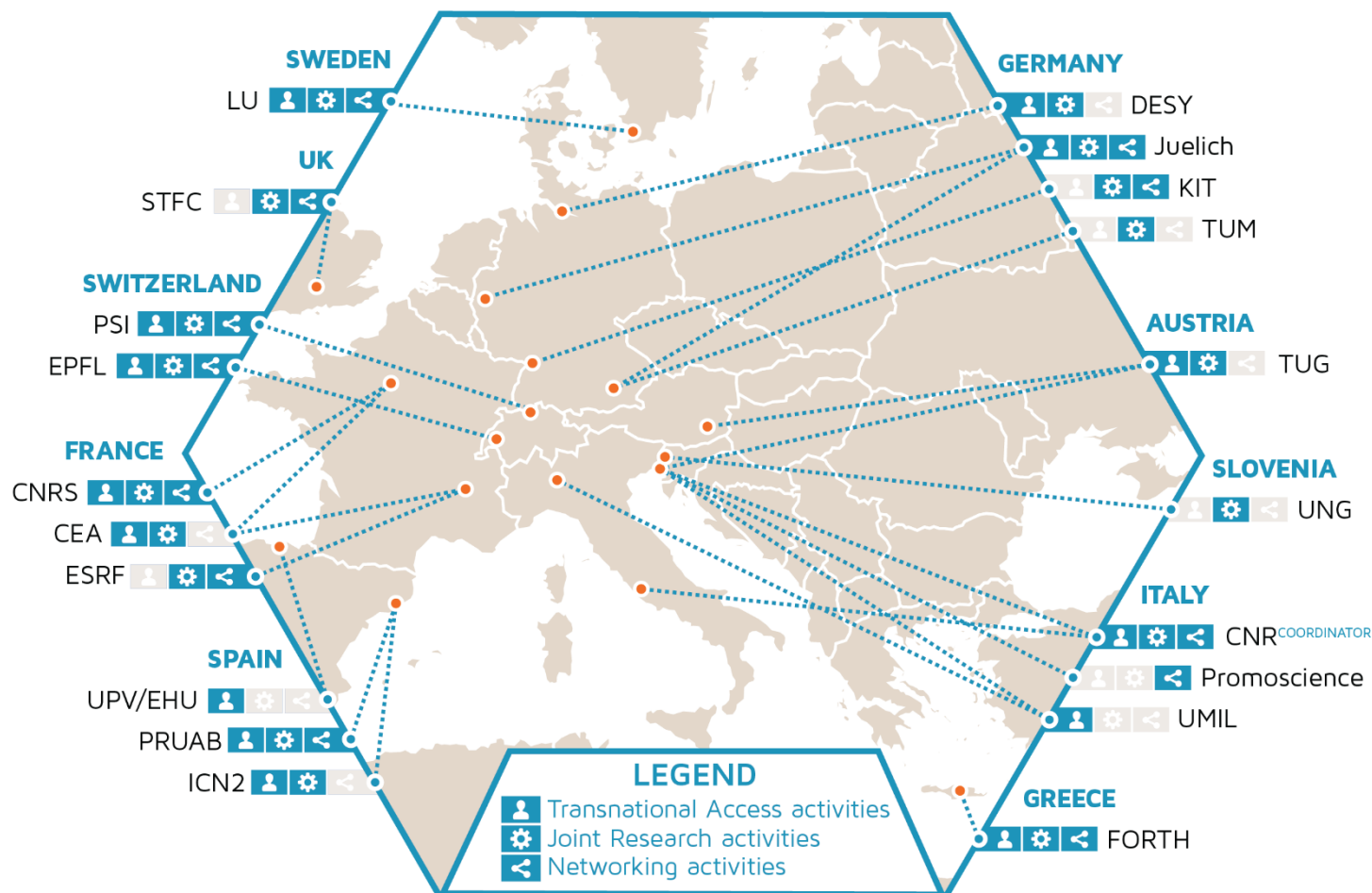


www.nffa.eu

- EU funded project
 - From 09.2015 to 08.2019
- it provides the widest range of tools for research at the nanoscale
 - Free transnational access to academia & industry
- Led by CNR-IOM

The consortium

20 partners of which **10 nanofoundries** co-located with Analytical Large Scale facilities



The offer

TA

Transnational Access activities

Multidisciplinary research at the nanoscale performed at nano-laboratories and ALSFs

Integration of theory & numerical analysis with advanced characterization

NA

Networking activities

Interface for different user communities

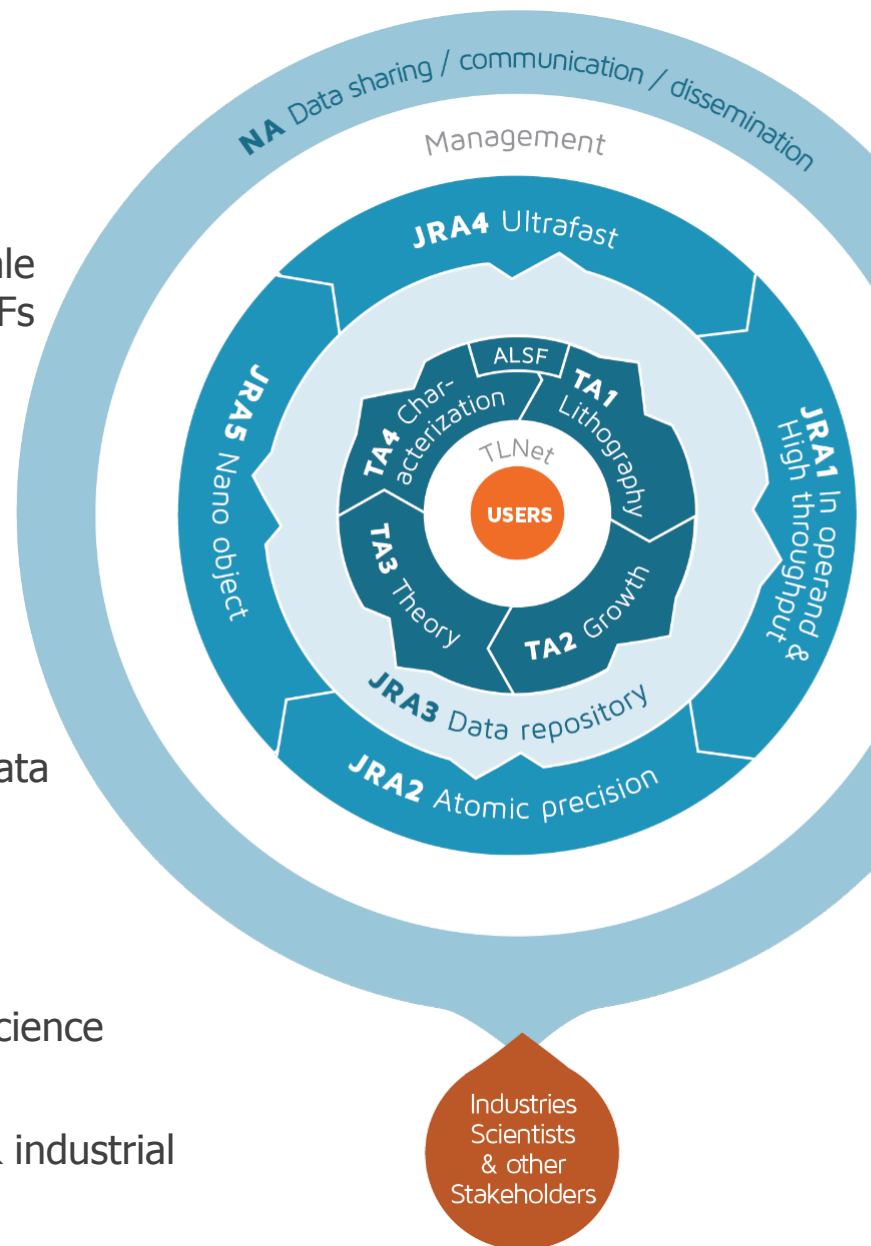
Industrial exploitation of experimental data

JRA

Joint Research activities

Methods & tools at the frontier in nanoscience research

Improved infrastructures for academic & industrial projects

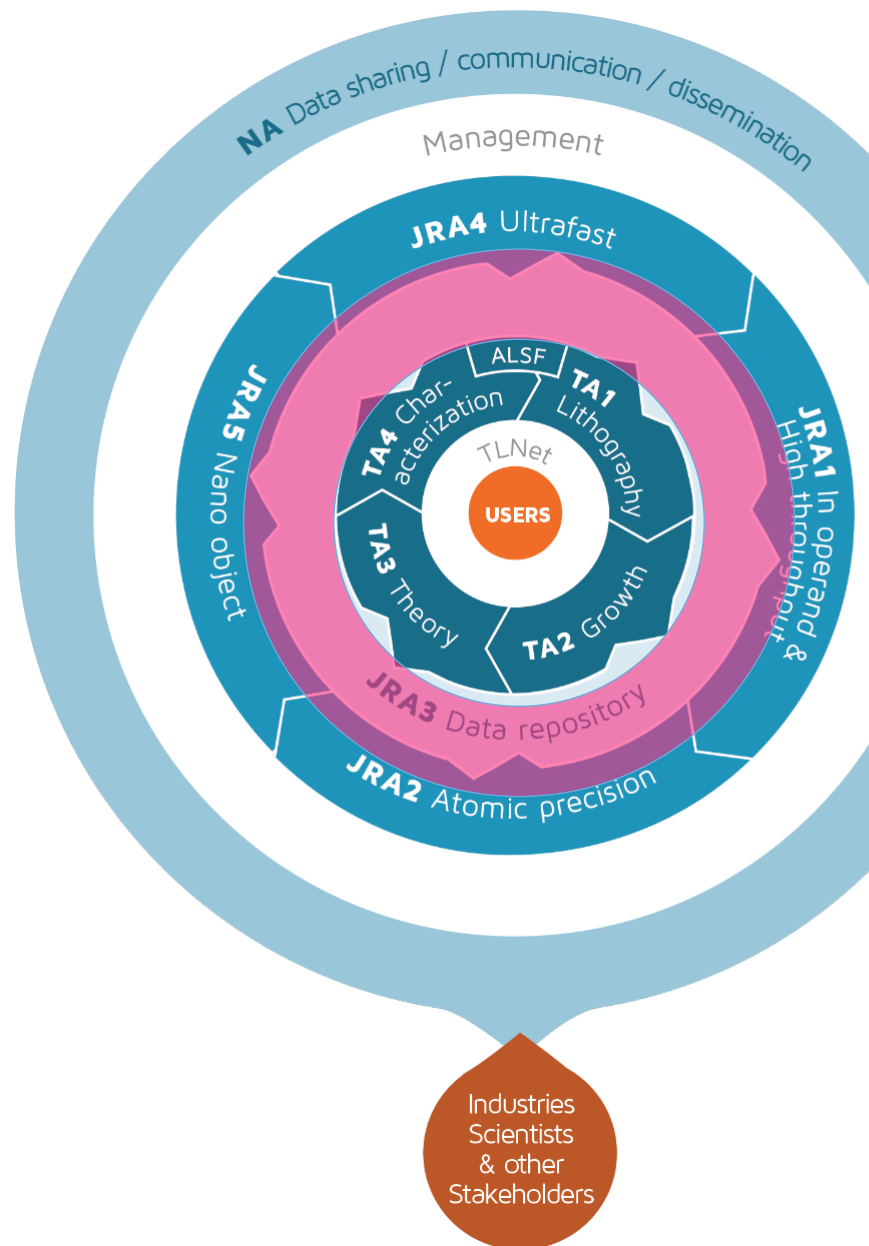


JRA on data...

JRA3

e-Infrastructure for data and information management

A transversal activity devoted to the setup of the first ***Information and Data Repository Platform (IDRP)*** for Nanoscience





The Mission

Manage and store the high volumes/varieties of scientific data generated at the NFFA facilities

Offer ***High Performance data access***

- Provide users with ***data sharing***, to validate findings, to facilitate reuse, and to promote collaborations and interoperability
- Identify and organize ***metadata*** associated to scientific data
- Make scientific data ***accessible and searchable*** by means of a metadata search engine



The importance of Metadata

The IDRP includes the relevant (searchable) information for:

- data analysis
- reproducibility of preparations and experiments

Scientific metadata

- citations, contacts, and collaborations

Basic metadata



JRA3 <-> NA collaboration Metadata Management

To develop metadata standards for the cataloguing, access and exchange of data and associated information describing nano-science experiments

- Draft metadata standard for nanoscience data released
 - Early deliverable as needed by WP8-JRA3 development
 - Developed a base metadata framework
 - distributed among international group interested in collaboration

Data Analytics and Management in Data Intensive D

XVIII International Conference, DAMDID/Ershovo, Moscow, Russia, October 11–14, Revised Selected Papers

Metadata for Experiments in Nanoscience Foundries

Vasily Bunakov¹✉, Tom Griffin¹, Brian Matthews¹, and Stefano Cozzini²

¹ Science and Technology Facilities Council, Harwell, Oxfordshire, UK
{vasily.bunakov,tom.griffin,brian.matthews}@stfc.ac.uk

² CNR, Istituto Officina dei Materiali, Trieste, Italy
cozzini@iom.cnr.it

Abstract. Metadata is a key aspect of data management. This paper describes the work of NFFA-EUROPE project on the design of a metadata standard for nanoscience, with a focus on data lifecycle and the needs of data practitioners who manage data resulted from nanoscience experiments. The methodology and the resulting high-level metadata model are presented. The paper explains and illustrates the principles of metadata design for data-intensive research. This is value to data management practitioners in all branches of research and technology that imply a so-called “visitor science” model where multiple researchers apply for a share of a certain resource on large facilities (instruments).

Keywords: Metadata · Nanostructures foundries

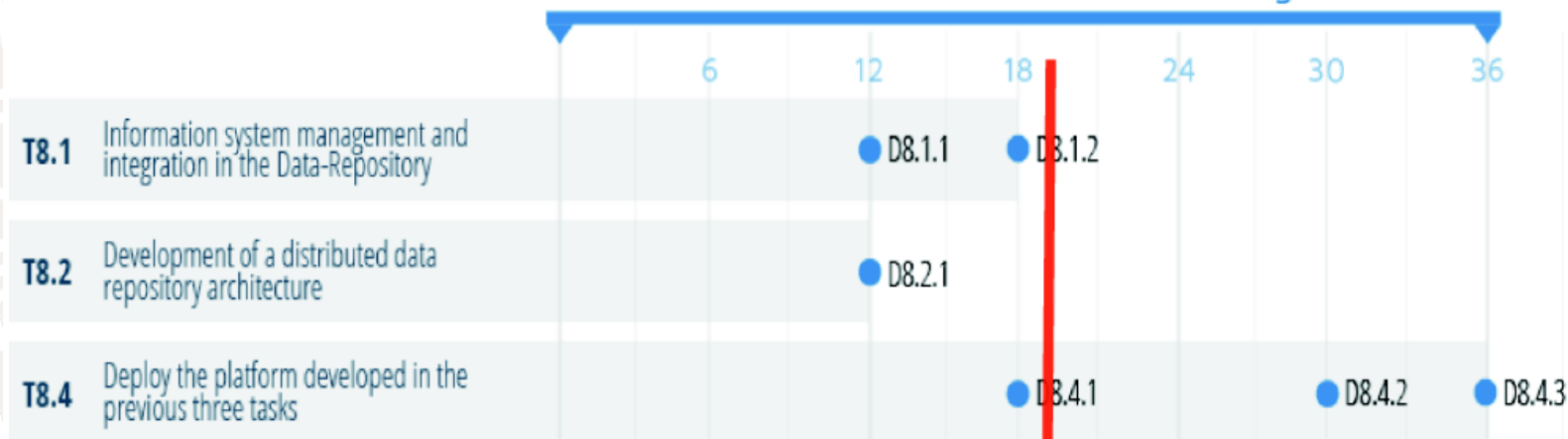
1 Introduction

The Nanostructures Foundries and Fine Analysis (NFFA-EUROPE) project (www.nffa.eu) brings together European nanoscience research laboratories that aim to provide researchers with seamless access to equipment and computation. This will offer a single entry point for research proposals, and a common platform to support the access and integration of the resulting experimental data. Both physical and computational experiments are in scope, with a vision that they complement each other and can be mixed in the same identifiable piece of research.

Metadata design is a part of a joint research activity within NFFA-EUROPE that takes empirical input from the project participants, and also takes into account state-of-the-art standards and practices. Metadata design is an incremental effort of the project;

Timeline

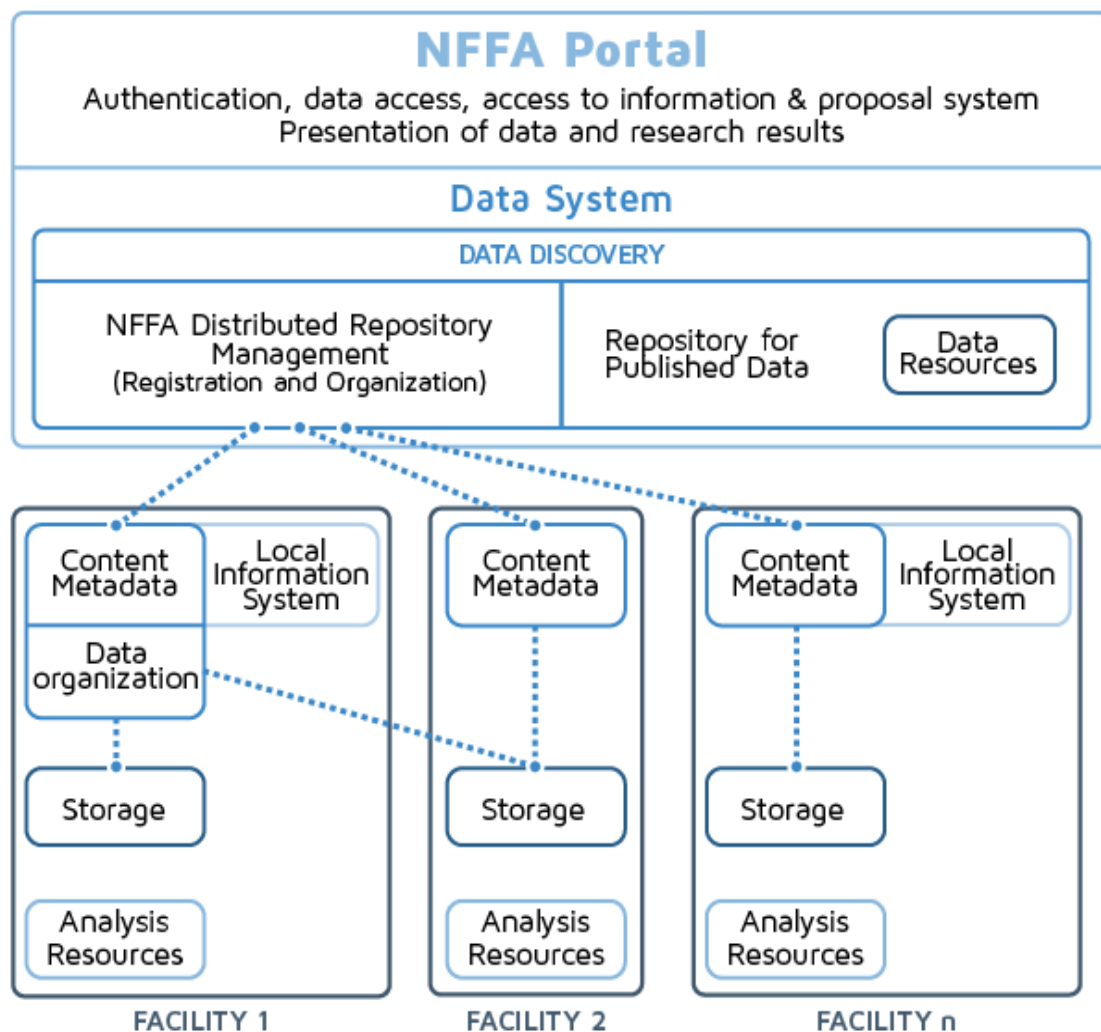
WP8 JRA3 - Research on e-infrastructure for data and information management



- D8.1.1 Design of Trusted authentication source for NFFA-Europe services [M12]
- D8.2.1 Design of the finalized repository architecture [M12]
- D8.1.2 Internal test of the information system [M18]
- D8.4.1 First Testbed available [M18]
- D8.4.2 Testbed fully deployed, including DASS services [M30]
- D8.4.3 Internal test of NFFA Distributed Repository [M36]

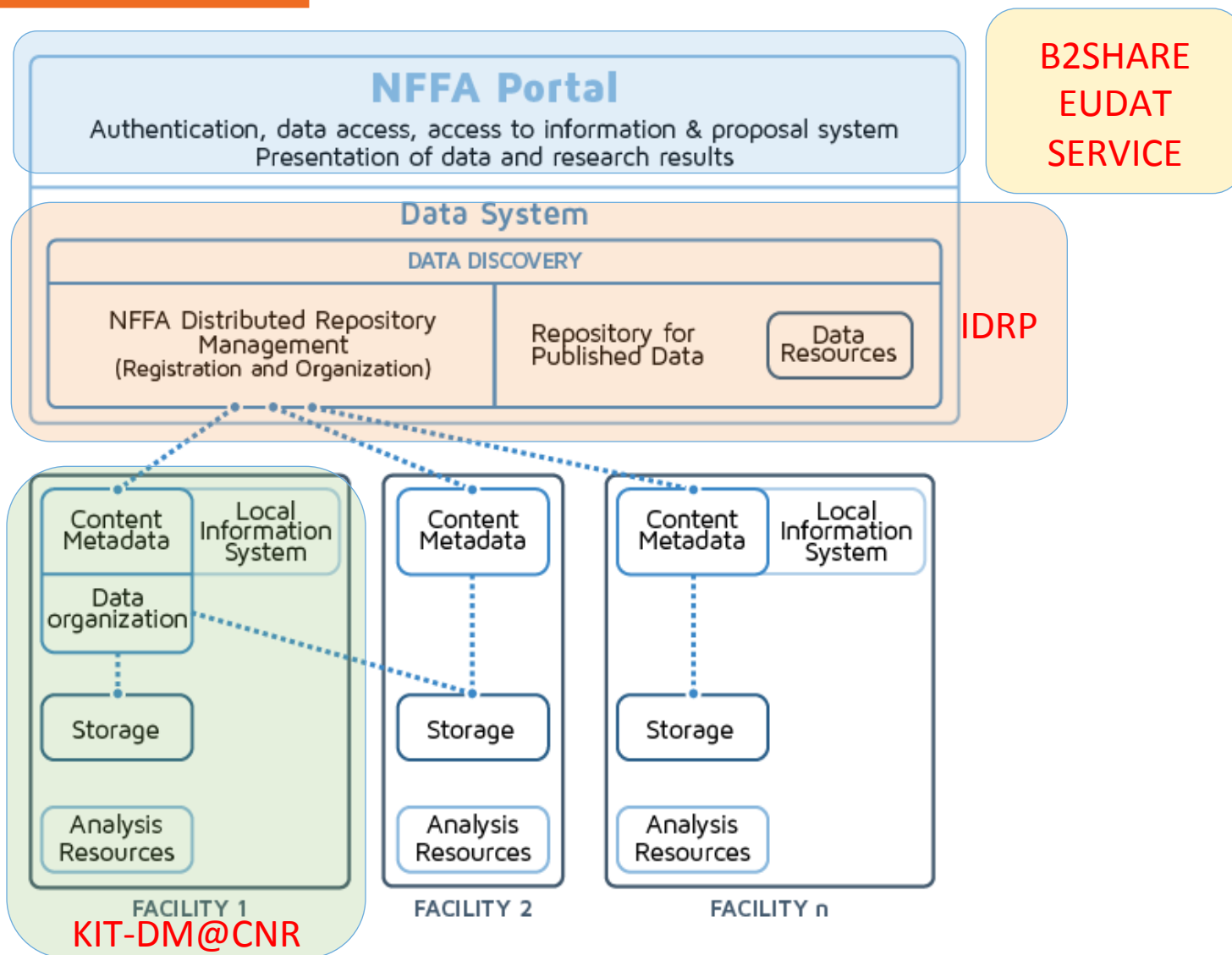


NFFA IDRP architecture

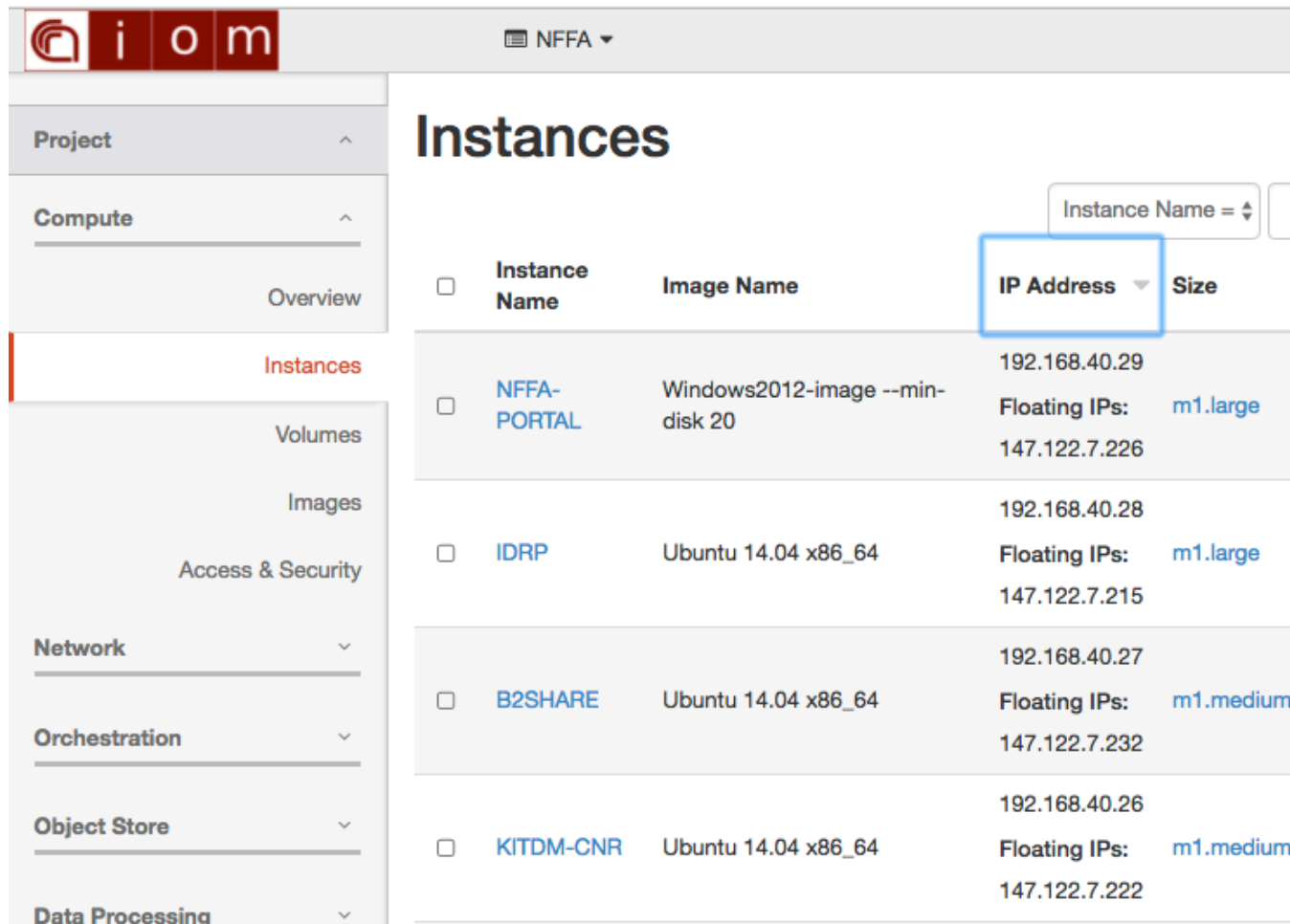




NFFA IDRP prototype



The IDRP prototype deployment...



The screenshot shows the 'iom' dashboard for the 'NFFA' project. The left sidebar contains navigation links: Project, Compute (selected), Overview, Instances (highlighted in red), Volumes, Images, Access & Security, Network, Orchestration, Object Store, and Data Processing. The main area is titled 'Instances' and displays a table of running instances. A blue box highlights the 'IP Address' column header. The table lists five instances: NFFA-PORTAL, IDRP, B2SHARE, and KITDM-CNR, each with its image name, IP address, and size.

Instance Name	Image Name	IP Address	Size
<input type="checkbox"/> NFFA-PORTAL	Windows2012-image --min-disk 20	192.168.40.29 Floating IPs: 147.122.7.226	m1.large
<input type="checkbox"/> IDRP	Ubuntu 14.04 x86_64	192.168.40.28 Floating IPs: 147.122.7.215	m1.large
<input type="checkbox"/> B2SHARE	Ubuntu 14.04 x86_64	192.168.40.27 Floating IPs: 147.122.7.232	m1.medium
<input type="checkbox"/> KITDM-CNR	Ubuntu 14.04 x86_64	192.168.40.26 Floating IPs: 147.122.7.222	m1.medium

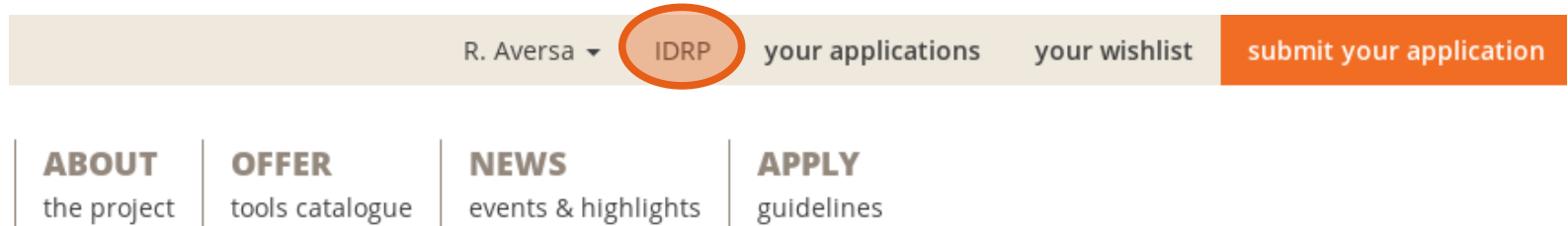


The prototype: NFFA Portal

A modified version of the current NFFA portal, in order to dialog with the IDRП.

It provides:

- Authorization and authentication services
- Information on the basic metadata related to the proposals submitted by NFFA users





The prototype: IDRP

The core of the architecture, connected both to the Data Management System and to the NFFA portal.

Possible actions:

1. register, manage and retrieve metadata and data stored in the local repositories
2. manage authorization of metadata and data access

Information and Data Repository Platform

MY PROPOSALS FIND MORE...

HOME

6 of 6 Proposals

🔍 (no filter active)

⬇ Proposal Id (ASC) ▼

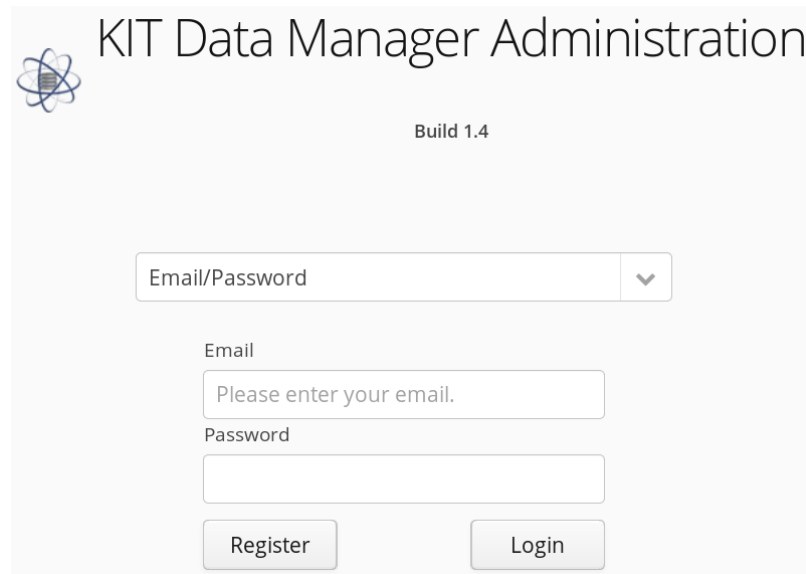
Proposal Details (FULL_ACCESS)

Proposal bc154604-ca44-458b-998a-e7771a304967

The prototype: KITDM@CNR

The Data Management System at the Local Facility adopted at CNR-IOM

Instruments data is ingested/downloaded to/from the KITDM@CNR, which creates the hierarchical metadata structure needed to store them



KIT Data Manager Administration

Build 1.4

Email/Password

Email

Password



What can you do with this?

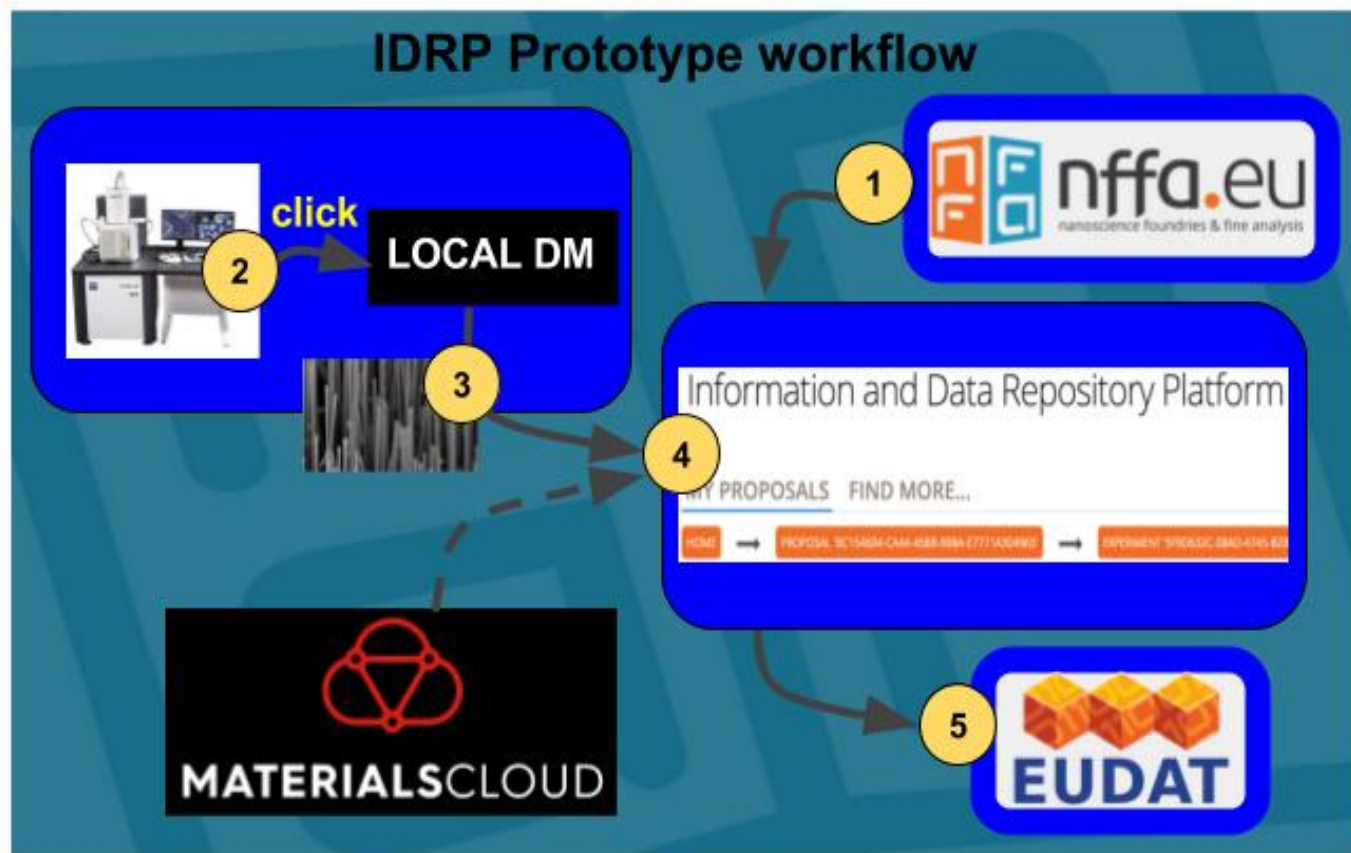
- Register/access using one single account
- Submit proposals **NFFA Portal**

- Ingest/download data generated at the facility to/from the Local Data Management Service **KITDM@CNR**

- Register data/metadata permanently to the IDRP
- Publish data/metadata to the IDRP or to B2Share at any moment
- Keep your data/metadata private for an embargo period
- Allow collaborators to access your data/metadata while they are under embargo

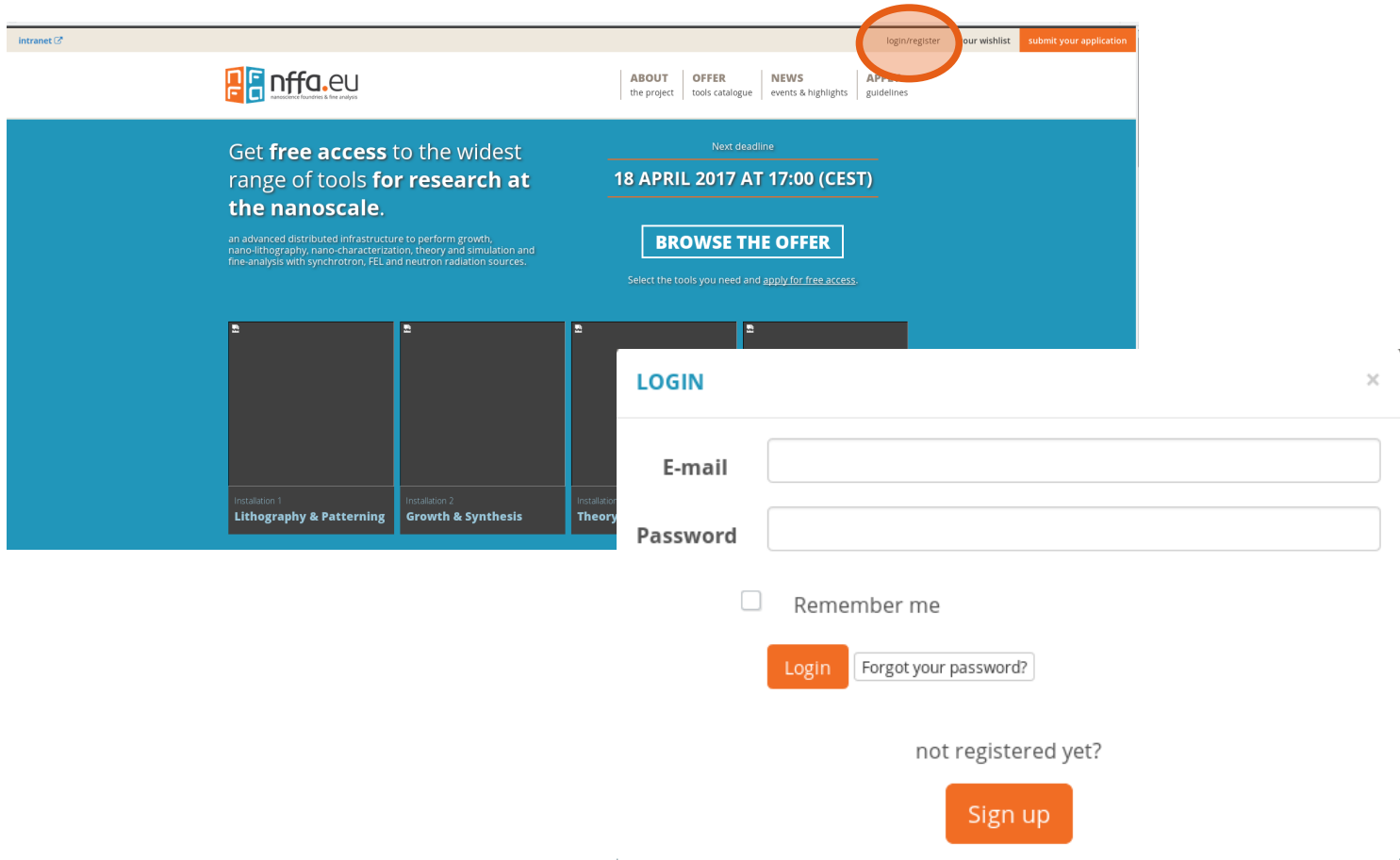
IDRP

A typical workflow



Basic use case: step 1

The user registers/authenticates on the NFFA portal



The screenshot displays the NFFA portal interface. At the top, a navigation bar includes links for 'login/register', 'your wishlist', and 'submit your application'. The 'login/register' link is circled in orange. Below the navigation bar, the main content area features a large blue banner with the text: 'Get **free access** to the widest range of tools **for research at the nanoscale.**' and a 'Next deadline' of '18 APRIL 2017 AT 17:00 (CEST)'. A 'BROWSE THE OFFER' button is prominently displayed. Below the banner, there are three installation cards: 'Installation 1: Lithography & Patterning', 'Installation 2: Growth & Synthesis', and 'Installation 3: Theory'. A 'LOGIN' modal window is overlaid on the right side of the page. It contains fields for 'E-mail' and 'Password', a 'Remember me' checkbox, and 'Login' and 'Forgot your password?' buttons. At the bottom of the modal, there is a link for 'not registered yet?' and a 'Sign up' button.

Intranet

login/register your wishlist submit your application

nffa.eu
nanoscale factories & fine analysis

ABOUT the project OFFER tools catalogue NEWS events & highlights APPLICATIONS guidelines

Get **free access** to the widest range of tools **for research at the nanoscale.**

Next deadline
18 APRIL 2017 AT 17:00 (CEST)

BROWSE THE OFFER

Select the tools you need and [apply for free access](#).

Installation 1
Lithography & Patterning

Installation 2
Growth & Synthesis

Installation 3
Theory

LOGIN

E-mail

Password

☐ Remember me

Login Forgot your password?

not registered yet?

Sign up

Basic use case: step 2

The user submit a proposal via the NFFA portal, providing contents for a basic metadata set:

- Associated project
- Requested facilities
- Samples to study

new application

Please read carefully the [guidelines](#) for the usage of the Single Entry Point and for proposal submission

NFFA-Europe very much welcomes and encourages all projects supporting innovation and involving industrial partners. NFFA recommends the main proposers to highlight in an appropriate way the involvement of an industrial partner, and we encourage eventual industrial partners to join as co-proposers if relevant.

ID

-

the ID will be assigned at the first "save draft"

your applications

NEW APPLICATION

FILTER PROPOSALS

Submitted

☐ HIDDEN

377

Proposal Title

Submitted

submitted on 02/03/2017 - 11:15

PREVIEW

proposal

title *

ERC sectors *

sector codes separated by a comma

download  the sector list and complete the field with the selected codes, separated by a comma

keywords *

min. 3 keywords separated by a comma

abstract *

min 1000, max 2000 characters (spaces included)

characters: 0 of 2000 | remaining: 2000

state of the art *

min 1500, max 3000 characters (spaces included)



Basic use case: step 3

The proposal basic metadata are ***automatically*** registered to the IDRP

Information and Data Repository Platform

MY PROPOSALS FIND MORE...

HOME

6 of 6 Proposals

▼ (no filter active)

Proposal Id (ASC)

Proposal bc154604-ca44-...

★ SWAGGER

🔊 47 Experiment(s)

Proposal bc154604-ca44-...

★ SWAGGER

🔊 47 Experiment(s)

Proposal bc154604-ca44-...

★ SWAGGER

🔊 47 Experiment(s)

Proposal bc154604-ca44-...

★ SWAGGER

🔊 47 Experiment(s)

Proposal bc154604-ca44-...

★ SWAGGER

🔊 47 Experiment(s)

NEW EXPERIMENT

DIRECT LINK

Proposal Details (FULL_ACCESS)

Proposal bc154604-ca44-458b-998a-e7771a304967

🔊 This a test by Promoscience

SWAGGER (PI)

EDIT

DETAILS

9 of 9 Experiments in Proposal

▼ (no filter active)

Experiment Id (ASC)

Placeholder Experiment using instrument 2045@CEA-LETI

🕒 16/02/2017, 08:23:44 - 16/02/2017, 08:23:44

🔊 1 Measurement(s)

🔊 One description

EDIT

DETAILS

NEW MEASUREMENT

Placeholder Experiment using instrument 1842@CNRS

🕒 16/02/2017, 08:23:39 - 16/02/2017, 08:23:39

🔊 1 Measurement(s)

🔊 No description

EDIT

DETAILS

NEW MEASUREMENT

NEW EXPERIMENT



Basic use case: step 4

- Measurements are performed at the facility
- When a dataset is uploaded, the corresponding metadata are automatically registered on the IDRP, as Data Asset with an associated Stream Identifier (location on the KITDM@CNR)

```
rossella@Barbie:~/CLI-KITDM$ ./click upload file --local A66C10.hdf5 --digital_object 57
scheduled ingest for digital object 57 (test_file)
uploaded file 'A66C10.hdf5' to '/webdav/admin/ingest_57_0733e613/data/A66C10.hdf5'
closed ingest for digital object 57 (test_file)

Trying to register on the IDRP...

post values {'dataAssetName': 'Data asset corresponding to test_file',
             'assetType': 'PLAIN_DATA', 'dataArchiveId': 'generic_archive_1',
             'dataStreamIdentifier': 'http://KITDM/rest/organization/download/57/',
             'dateOfCollection': '2017-08-23T14:39:52Z',
             'measurementId': 'ecac3fb6-ed81-4f70-acfc-87eddb573e1d'}

Data Asset registered on the IDRP!
```



Basic use case: step 5

The user can also manually register data and/or metadata on the IDPR

Data Asset Location

Basic Asset Information

Basic Asset Information

Data Asset Overview

Summary

At this point, basic data asset information are captured. These information are:

- Data Asset Name: An possibly meaningful name of the data asset, e.g. a file name. The provided name will be stored in lowercase and characters forbidden to be used in filenames will be escaped.
- Data Stream: The URL pointing to the data stream. This URL must be chose according to the previously selected data archive. Typically, this should be an HTTP URL.
- Date of Collection: The date when this data asset has been collected. By default, the current date is used.

Optionally, you can check whether the provided Data Stream can be accessed by the IDRP using the 'Check Accessibility' button. If this is not the case, transparent accessfor the user won't be possible. Instead, the user will be provided with contact information if (s)he requests a download of the particular data asset.

asset type

LANDING_PAGE (External landing page accessible by the Web browser.)

asset stream or access identifier *

http://materialscloud.org/archive/2017.0003/v1/

asset name *

Landing Page 2017.0003

date of collection

22/08/2017

CHECK ACCESSIBILITY

BACK

NEXT



Some comments..

It is a ***prototype***: many features can be added and/or enhanced and/or modified according to the community-specific needs.

It is ***modular***: the whole platform or single services independently can be adopted

Distributed Repository achievements:

- 1) Identification and organization of metadata
- 2) Sharability of scientific data
- 3) Scientific data policy
- 4) Authentication service

A case study: classifying SEM images by Neural network..

- We created and manually annotated the first sample of classified SEM images (for a total of 18,577 images).
- The classified SEM image are stored on the KITDM as a standalone set

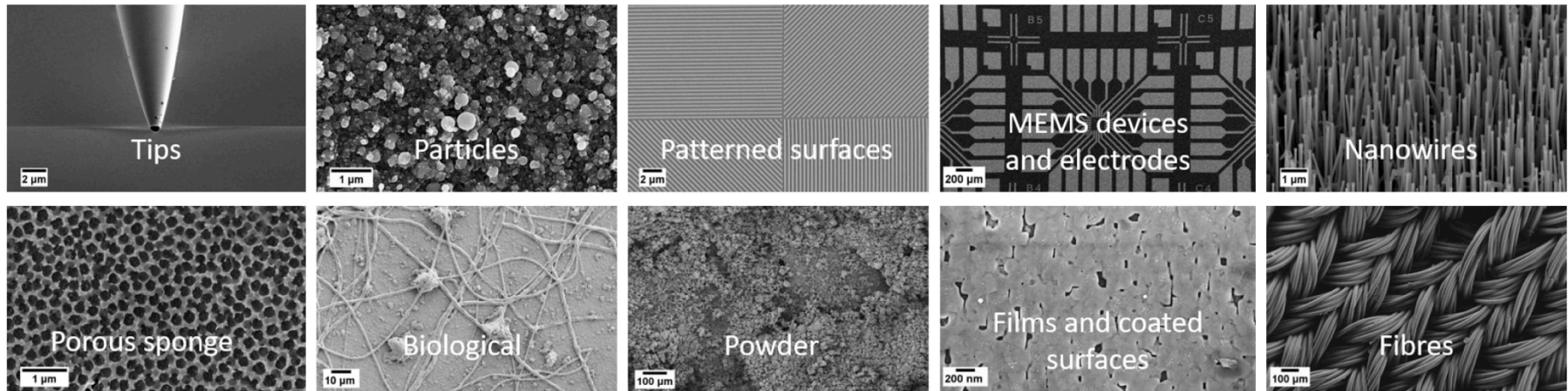



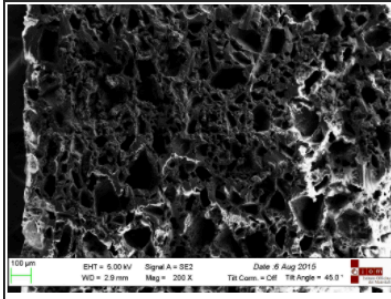
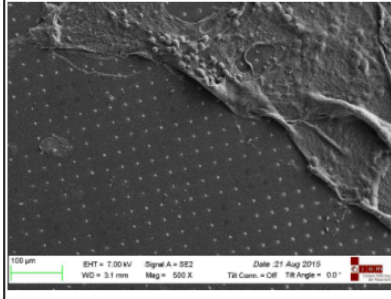
Figure 1. Categories chosen for SEM images. The dimensionality of nanoscience objects provided the basis for the choice. Other categories, such as Biological and Tips were added as these were common images found in the SEM database.



A case study: classifying SEM images by Neural network (2)

- We applied the deep learning technique on the above dataset to train an automatic image classification engine
 - the average test of the algorithm achieved 95% accuracy..
- A full automated procedure has been setup to automatically annotate SEM images once user load them on the KITDM&IDRP..
 - Metadata generated automatically thanks to ML

Our SEM image web classifier

Image	Processing page	Nanowires	MEMS devices electrodes	Biological	Films Coated Surface	Porous Sponge	Patterned surface
	237	0.2% <input type="radio"/>	4.1% <input type="radio"/>	0.03% <input type="radio"/>	0.9% <input type="radio"/>	93.6% <input checked="" type="radio"/> Predicted	0.31% <input type="radio"/>
	238	00.07% <input type="radio"/>	0.00% <input type="radio"/>	99.9% <input checked="" type="radio"/> Predicted	0.00% <input type="radio"/>	0.00% <input type="radio"/>	0.00% <input type="radio"/>

Submit changes



The NFFA-IDRP data policy



A project funded by the European Union

www.nffa.eu

secretariat@nffa.eu

NFFA-EUROPE IDRP

SCIENTIFIC DATA POLICY

Draft version 1.2 date 20.03.2017.

- Prepared and discussed internally by JRA3
- Presented/Discussed at Management Team
- Presented and discussed at the Executive committee
- Approved by the General Assembly

Why such a document ?

- NFFA-Europe IDRP defines a policy on data in order to:
 - Ensure the continuing availability of data of long-term value for research, teaching, and for wider exploitation by individuals, government, business, or other organizations;
 - Support the integrity, transparency and openness of the research;
 - Help in the formal publication of data sets, as well as enable the tracking of their usage through citation and data licenses



Structure of the document

- Short Preamble
- General principles
- Raw data and associated metadata
 - Treatment of raw data and associated metadata
 - Access to raw data and metadata
- Good practice for metadata capture and results storage
- Appendix 1: Definitions of terms

1 General principles

1. The present data management policy pertains to the ownership of, the curation of, and access to experimental data and metadata collected and/or stored on the NFFA-Europe IDRPs within the context of the NFFA-Europe project.
2. Acceptance of this policy is a condition for the award of NFFA-Europe IDRPs facility usage.
3. Users must not attempt to access, exploit, or distribute raw data or metadata unless they are entitled to do so under the terms of this policy.
4. Deliberate infringements of the policy may lead to denial access to raw data or metadata, and/or denial of future facility usage requests within the NFFA-Europe.
5. All data and metadata will be subject to the data protection legislation of the Countries in which the data and metadata are stored.



Summary & Results

- NFFA IDRП has been designed and deployed..
- IDRП prototype up & running and under testing
 - Six different groups involved..
 - Positive feedback received
- Significant scientific case study evaluated:
 - Data analysis services under development
- Data Policy document prepared, discussed and approved
 - NFFA users have a clear picture..
- Cooperation with EU projects and RDA established and fruitful
 - Metadata collaboration with RDA
 - Eudat services implemented

Thank you

CONTACTS

stefano.cozzini@iom.cnr.it

+39 040 3787508

www.nffa.eu



Backup slides



International interaction (1)



EUDAT

SERVICES & SUPPORT ▾

COMMUNITIES & PILOTS

WORKING GROUPS ▾

NFFA-EUROPE Information and Data Management Repository Platform for nano-science in Europe Data Pilot



An insight into the NFFA-EUROPE Information and Data Management Repository Platform for nano-science in Europe Data Pilot

Attachment:

 [PilotProfile-EUDAT-NFFA.ppt](#)



International interaction (2)

Strict collaboration with the forthcoming EU project JRA3 technology/expertise will contribute in the data repository

WELCOME TO E(U)SMI // THE EUROPEAN SOFTMATTER INFRASTRUCTURE

ESMI turns EUSMI

There is good news for the European Soft Matter community:

The Grant Agreement for EUSMI, the designated successor project of ESMI was signed on 30 April 2017.

Therefore, we will be able to provide the full service, which the community was used to under ESMI, starting from

1. July 2107.

2 Raw Data & associated metadata

2.1. Treatment of raw data and associated metadata

2.1.1. Basic metadata referring to the accepted proposal will be made available on the NFFA-Europe IDRPP portal at different stages. By default, the IDRPP publishes the following basic metadata: PI, title of proposal and instruments to be used are released once the proposal is accepted. Once the embargo period is completed abstract of the proposal is released.

2.1.2. Associated/contextual metadata and links to the data obtained as a result of publically funded access to the NFFA-Europe research facilities will be released open access after an initial embargo period during which access is restricted to the research users, represented by the PI. The embargo period can be extended according to the section 2.2 of this policy.

2.1.3 Industrial users may opt for a proprietary access where all data and metadata remain confidential [...]

2.1.4. Raw data stored in facilities that are part of the NFFA-Europe are under the custody of these facilities, while NFFA-Europe is the custodian of all of associated metadata stored on the NFFA-Europe IDRPP [...]

2.1.5. The NFFA-Europe plans for 10 years as long-term period to maintain metadata and potentially data within the IDRPP [...]

2 Raw Data & associated metadata

2.1. Treatment of raw data and associated metadata

2.1.1. Basic metadata referring to the accepted proposal will be made available on the NFFA-Europe IDRPP portal at different stages. By default, the IDRPP publishes the following basic metadata: PI, title of proposal and instruments to be used are released once the proposal is accepted. Once the embargo period is completed abstract of the proposal is released.

2.1.2. Associated/contextual metadata and links to the data obtained as a result of publically funded access to the NFFA-Europe research facilities will be released open access after an initial embargo period during which access is restricted to the research users, represented by the PI. The embargo period can be extended according to the section 2.2 of this policy.

2.1.3 Industrial users may opt for a proprietary access where all data and metadata remain confidential [...]

2.1.4. Raw data stored in facilities that are part of the NFFA-Europe are under the custody of these facilities, while NFFA-Europe is the custodian of all of associated metadata stored on the NFFA-Europe IDRPP [...]

2.1.5. The NFFA-Europe plans for 10 years as long-term period to maintain metadata and potentially data within the IDRPP [...]

3 Good practice for metadata capture and results storage

3.1. The research users are encouraged to ensure that the experiments metadata are as complete as possible, as this will enhance the possibilities for everybody to search for, retrieve and interpret the data in the long term.

3.2. NFFA-Europe provides means for the capture of such metadata items that are not automatically captured by the instrument, in order to facilitate recording the fullest possible description of the raw data.

3.3. Data users who aim to carry out analyses of raw data and metadata, which are openly accessible, are invited to contact the original PI or his/her designate to inform them him/her and suggest a collaboration, if appropriate. Researchers must acknowledge the source of the data and cite its unique identifier as well as any publications linked to the same raw data.

3.4. PIs and user group members who carry out analyses of raw data are encouraged to link the results of these analyses to the raw data and metadata. Furthermore, they are encouraged to make such results openly accessible.

3.5. PIs and user group members who carry out analyses are encouraged to publish well defined datasets, that can include both raw and analyzed datasets, associating persistent identifiers with them, in order to make data identifiable and citable.

3.6. Publications related to data from experiments carried out at NFFA-Europe facilities are invited to cite the proposal ID and, if available, persistent identifiers of the experiment and data.

4 Appendix: definition of terms

- 4.1. The term raw data pertains to data collected from experiments performed on NFFA-Europe instruments, including results of computer experiments (simulations) and uploaded to the NFFA-Europe IDRP. This definition includes data that are created automatically or manually by facility specific software and/or facility staff expertise in order to facilitate subsequent analysis of the experimental data.
- • 4.2. The term metadata describes information pertaining to research proposals submitted via NFFA portal (which may be referred to as basic metadata), as well as data collected from NFFA-Europe facility instruments, including (but not limited to) the context of the experiment, the research users, experimental conditions, and other logistical information stored on the NFFA-Europe IDRP (which may be referred to as associated metadata).
- • 4.3. The term principal investigator (PI) pertains to the main proposer identified on the NFFA-Europe proposal as “user group leader”.
- • 4.4. The term research user refers to any person of the proposal research team or to whom the PI designates the right to access resultant raw data and associated metadata.
- • 4.5. The term data users refers to anyone accessing the IDRP portal and metadata therein.
- • 4.6. The term public research refers to research done through peer review and leading to publication(s).
- • 4.7. The term online data catalogue pertains to a computer database of metadata containing links to raw data files, that can be accessed by a variety of methods, including (but not limited to) web-based browsers.
- • 4.8. The term results pertains to data, intellectual property, and outcomes arising from the analysis of raw data. This does not include publications.
- • 4.9. The term custodian refers to the Institute storing, curating, and providing access to raw data, metadata, and results.
- • 4.10. The term long-term means a minimum of 5 years up to 10 years.
- • 4.11. The term user groups members refers to the research users associated with a proposal. They are identified and associated to the proposal by the PI.
- • 4.12. The term open data access refers to the definition reported in the “Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020” mentioned above.
- • 4.13. The term SME stands for Small and medium-sized enterprises and are defined in the EU recommendation 2003/361.d, if available, persistent identifiers of the experiment and data.