nffa.eu
PILOT 2021 2026

DELIVERABLE REPORT

WP2 - MGT2 - Pilot scheme for the management of a distributed research infrastructure offering harmonised, interoperable and integrated services

D2.5

First assessment of the data management procedures

This initiative has received funding from the EU's H2020 framework program
for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

PROJECT DETAILS

PROJECT ACRONYM	PROJECT TITLE
NEP	Nanoscience Foundries and Fine Analysis - Europe|PILOT
GRANT AGREEMENT NO:	FUNDING SCHEME
101007417	RIA - Research and Innovation action
START DATE	
01/03/2021	

WORK PACKAGE DETAILS

WORK PACKAGE ID	WORK PACKAGE TITLE
WP2	MGT2 - Pilot scheme for the management of a distributed research infrastructure offering harmonised, interoperable and integrated services

WORK PACKAGE LEADER

Dr. Cristina Africh (CNR)

DELIVERABLE DETAILS

DELIVERABLE ID	DELIVERABLE TITLE
D2.5	First assessment of the data management procedures

DELIVERABLE DESCRIPTION

Effective implementation of data management procedures.

DUE DATE	ACTUAL SUBMISSION DATE
M30 (Month) 31/08/2023	30/11/2023

AUTHORS

Francesco de Giorgi (eXact lab)
Giuseppe Piero Brandino (eXact lab)

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

Andrea Recchia (eXact lab)
Rossella Aversa (KIT)
Mirco Panighel(CNR)
Irene Modolo (CNR)

PERSON RESPONSIBLE FOR THE DELIVERABLE

Francesco de Giorgi (eXact lab)

NATURE

☒ R - Report

□ P - Prototype

□ DEC - Websites, Patent filing, Press & media actions, Videos, etc

□ O - Other

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

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

REPORT DETAILS

ACTUAL SUBMISSION DATE	NUMBER OF PAGES
30/08/2023	**16** 16

FOR MORE INFO PLEASE CONTACT

Francesco de Giorgi
eXact lab
Via Crispi 56, 34126, Trieste, italy

email:francesco.degiorgi@exact-lab.it

VERSION	DATE	AUTHOR(S)	DESCRIPTION / REASON FOR MODIFICATION	STATUS
1	8/7/2023	Giuseppe Piero Brandino	Structure layout	Draft
2	11/23/2023	Francesco de Giorgi Giuseppe Piero Brandino Andrea Recchia Rossella Aversa Mirco Panighel Irene Modolo		Draft ready for review
				Choose an item.
				Choose an item.
				Choose an item.
				Choose an item.
				Choose an item.

This initiative has received funding from the EU’s H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

CONTENTS

This initiative has received funding from the EU’s H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

DELIVERABLE REPORT

This document contains the first report on the status of the data management in NFFA Europe Pilot. After a summary of the general philosophy used for the data management, we will briefly recall the solutions implemented, and we will present the initial statistics and/or feedback from partner and users.

Data Management approach in NEP

The general idea behind the data management procedures in NFFA Europe Pilot is the ease of use, without giving up versatility and extendability. This results in a fairly complicated logical and IT infrastructure, mostly hidden to the final user.

From the point of view of the user, the typical flow can be schematized as

1. Collect the data from the instrument and store them temporary (e.g. on a lab machine, or on a data collaboration platform like the NFFA Datashare [1])
2. Describe the data with commonly agreed metadata, corresponding to existing metadata schemas.
3. Automate the metadata compilation and mapping to the schema as much as possible:
   a. Extract metadata from the instrument during data generation and/or
   b. Extract metadata from the generated data
   c. Fill the missing metadata using the Metadata Editor [2]
4. Optional: Use a data converter to convert data and metadata to a common format (e.g. NeXus) following a NeXus application definition
5. Upload the data (or optionally the resulting NeXus file) on a data repository (e.g. Zenodo) and make it publicly available
6. Upload the metadata document created at point 3 on the NEP metadata repository MetaRepo [3], including the reference to the data uploaded to the data repository at point 5

Most of these steps can be carried on using services and tools created or indicated by NFFA Europe Pilot. The only steps that could require potentially a custom piece of software to be developed are step 3 and 4, i.e. the metadata extraction and mapping and the data conversion.

From the point of view of the logic behind and its implementation, the picture is fairly more complex, in particular regarding the various interactions among services and applications. A schematic representation is reported in figure 1.

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

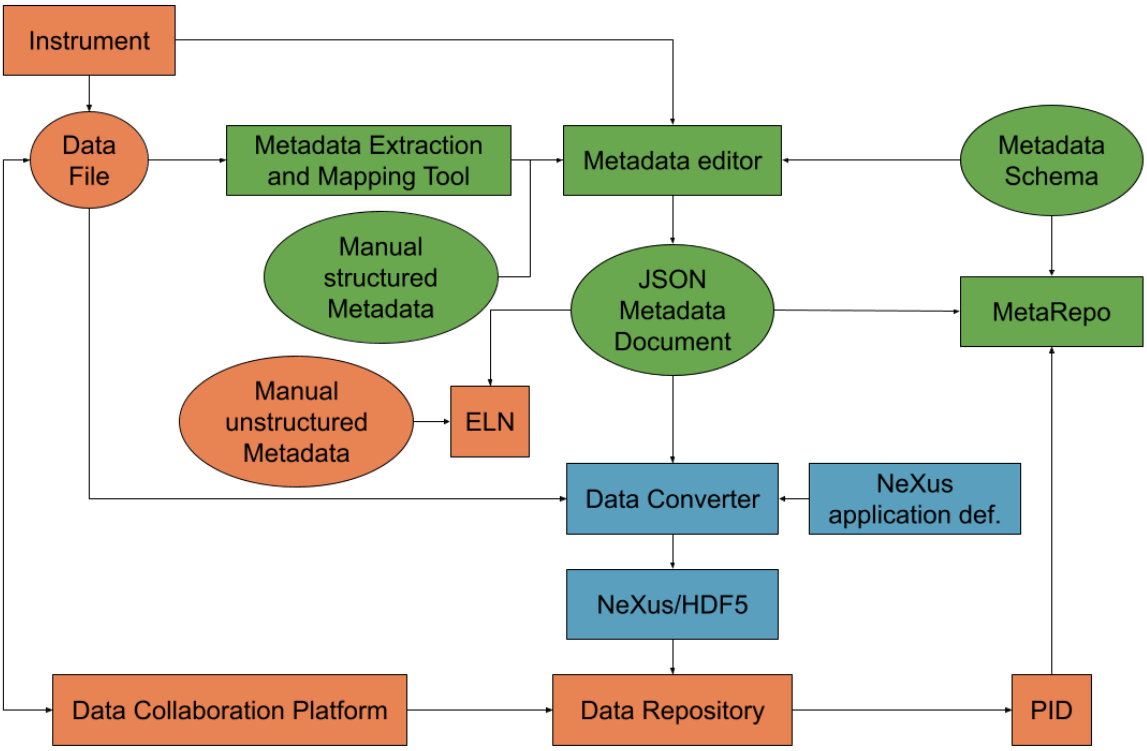


Figure 1: Logical representation of the data management infrastructure. Squares: Services and tools, data and metadata: circles. Orange: minimal scenario, green: optimal data management scenario, blue: optional advanced scenario.

Figure 1 details the various interactions between instruments, services, tools and repositories following the data/metadata lifecycle, indicated by arrows.

In detail: the data file generated by the instrument can be temporarily stored on a Data Collaboration Platform (point 1), and then published in a Data Repository (point 5) which assigns a persistent identifier (PID) to it. If the metadata schema is available for a given technique (point 2), the metadata can be automatically extracted from the instrument at data generation time by means of a dedicated API (point 3.a) or from the generated data file using the Metadata Extraction and Mapping Tool (point 3.b). In both cases, they can be fed into the Metadata Editor to be rendered and manually checked, completed or edited (point 3.c). The result is a JSON metadata document, which can be uploaded (and validated against the metadata schema) in the MetaRepo and related to the data it describes via the PID (point 6) using the MetaRepo API, the MetaRepo web GUI, or the Metadata Editor. The same metadata document can be imported into an ELN, to manually include any additional unstructured metadata, in case this is needed. Optionally, the data and the metadata document can be fed to a Data Converter following a NeXus application definition (equivalent to the metadata schema) to create a NeXus file which can be uploaded to the Data Repository.

Data management plan and data stewardship wizard

The Data Stewardship Wizard (DSW) is an open source tool that supports the creation of DMPs through a questionnaire, in which users are guided by resources, recommendations and definitions, that allows to structure the answers into a formatted document, the Lab-DMP.

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

A self-hosted DSW instance, specifically adapted to include the relevant data management practices for the NFFA-Europe Pilot project, was deployed to support researchers in creating the DMP for their laboratories. The instance can be reached at [4] and accessed using the NFFA-Europe credentials (see D16.1 and 16.2).

At present, about 48 DMPs have been generated by the laboratories of the infrastructure. The DMPs, beside recalling the obligations for a FAIR data management, provide information on data collection (file formats and analysis software used, naming conventions), storage (repositories and use of PIDs) and reuse (use of metadata schemas and ELN) for each laboratory.

Metadata schema development strategy

Many groups within the broad field of Materials Science are already developing metadata schemas and ontologies, driven by their individual organizational goals and guided by the open and FAIR data initiatives. To promote interoperability, a harmonization approach to find a common description of the data coming from the measurements needs to be developed and adopted as far as possible. Motivated by this, a coordination effort was established in 2021 between the NEP WP16 and the Joint Lab "Integrated Model and Data Driven Materials Characterization" (MDMC) of the Helmholtz Association.

The final objective of the incremental process we are undertaking is to describe the entire workflow of a research Study, providing rich metadata to make data interoperable. Priority has been given to experimental techniques. Figure 2 shows an example of a basic, simplified workflow of an experimental Study in which every process is performed only once. The user(s) who performed the processes is omitted in the graphics for the sake of simplicity.

As for our strategy, we decided: i) to develop the metadata schemas in JSON; ii) to adopt terms from existing schemas or ontologies, when relevant, and iii) to adhere to the NeXus [5] naming convention as much as possible.

We created a modular solution, in which every output (minimal: sample, raw data, analyzed data; optional: input, precursor, sample component, processed data) is described by a metadata document according to a schema. The schema of the output contains the description of the process(es) needed to obtain it and the reference of the used input(s) and to other related resources. By reference, we mean the URI to the metadata document in MetaRepo, whenever available. The data resource is linked using the relatedResource field in the administrative metadata of the metadata document record. With this approach, there are no mandatory steps or inputs in the workflow. The only important thing is to relate a given output to the necessary input(s) to reproduce it.

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

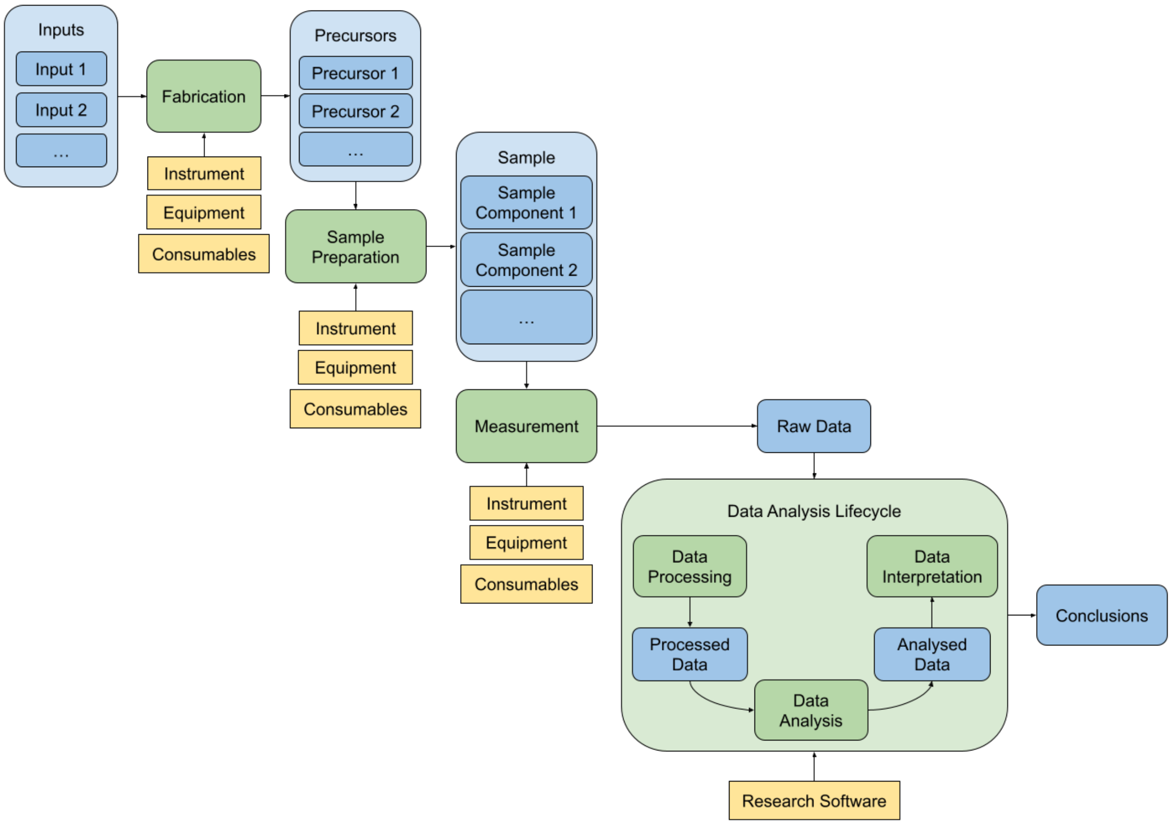


Figure 2: Basic experimental workflow. The steps are in green, the inputs and output of each step are in blue, the tools needed to perform each step are in yellow.

Data management tools

As seen in the previous section, the NFFA Europe Pilot data management procedures are broken down in individual steps, each of which correspond to a tool or service. A summary and, where available, some usage statistics and feedback are reported for the relevant part of the infrastructure.

Metarepo

MetaRepo, accessible at [3], is a generic metadata repository and metadata schema registry, based on MetaStore [6], developed and hosted at KIT. It enables data curators to register an arbitrary number of metadata schemas, and it allows research users to store metadata documents which are automatically validated at upload time against the corresponding metadata schema. The service is fully integrated into the NEP infrastructure [7] and has been available as Virtual Access since Month 25, with additional features included in the next version since Month 31 [8].

To date, including all the versions, MetaRepo currently contains 34 metadata schemas (16 unique), which are publicly available, and 151 metadata documents, which are accessible according to the individual access control lists. Total count of actions performed: 6713 number of users: 22

As examples of application use cases, MetaRepo has been employed to register metadata schemas and to store metadata documents in order to improve the FAIR data management of Magnetic

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

Resonance Imaging (MRI) data [9] and to represent Scanning Electron Microscopy (SEM) images as FAIR Digital Objects to compose training datasets for Machine Learning [10].

Metadata editor

Metadata editor is a user-friendly desktop application designed for managing metadata documents. With an intuitive interface, it facilitates the compilation and editing of metadata documents using metadata schemas stored on MetaRepo. The application handles validation, enables the export and upload of documents to MetaRepo, and provides functionality to reconstruct document provenance.

Guidance and documentation for the metadata editor can be accessed at [2]. The latest version of the application is available in the 'Download' section.

A more detailed description of the tool is available in deliverable 16.4 [11]. Furthermore, the tool, along with its placement within the data management workflow, has been presented in the "FAIR and Open Data in NFFA-Europe Pilot and Beyond" workshop.

Data repository

Any Research Data generated in NFFA Europe Pilot, including associated Metadata, needed to validate the Results presented in a Scientific Publication or appearing in it must be published on a publicly accessible and persistent Data Repository, as provided by the H2020 Open Research Data Pilot and in Chapter 5.2 of NFFA-Europe Research Data Policy [12].

NFFA Europe Pilot infrastructure does not include its own Data Repository for long term persistence, but at point 5.2.4 of NFFA-Europe Research Data Policy it requires its Beneficiaries and Research Users to upload their Research Data to an appropriate OpenAIRE compatible Data Repository [12]. There are a few reasons behind this choice: firstly, several public repositories are now mature and reliable projects, and developing or even just deploying a new one would have increased the complexity and efforts of the NFFA Europe Pilot IT infrastructure without providing benefits, with the risk of an underperforming service. Secondly, it would have been extremely complex, on the administrative and financial side, to guarantee the long term persistence of the Data Repository itself in the long run. Finally, choosing a platform already compatible with OpenAIRE, i.e. the EC funded infrastructure dedicated to the verification of Open Access policies, allows an effective monitoring of the project’s research outcomes both by the Project Management Axis and by the official bodies; as a matter of fact, NFFA-Europe Pilot project outcomes and basic statistics can be browsed and downloaded as a report at [13].

The repository of choice can be a discipline-specific Data Repository, an institutional one, or a multi-disciplinary open repository like Zenodo.

At the current moment, 8 open datasets related to NEP Scientific Publications are available, and all of them were published on Zenodo.

For the project's public reports, a Zenodo community has been created, available at [14]. It is managed by the Research Data management Team of NEP and contains all the public deliverables of the project, which, in this way, can be easily browsed and cited in other outcomes.

Data collaboration platform (NFFA Datashare)

NFFA-Europe makes available to its users the NFFA Datashare [1], a file sharing and collaboration platform based on Nextcloud open source software and hosted on servers under NFFA custody. It is accessible to all registered NFFA-Europe users.

Users can store, access and share Raw Data and Analyzed Data in every file format, collaborating in real time with our experts and other team members.

On a voluntary basis, NFFA-Europe Laboratories can install Nextcloud desktop client on each computer of the Laboratory that acquires or contains NFFA research data and synchronize the local folders containing NFFA research data to an account on NFFA Datashare. Then, the Head of the Laboratory shares to the personal NFFA Datashare accounts of the users the specific subfolder related to their experiment so that all the data produced and collected during the access are synced in real time to their personal NFFA Datashare accounts, without the risk of losing data by placing them on portable data storage devices.

NFFA Datashare currently hosts 1,580 users. Currently, 20 Laboratories, almost all located at the CNR-IOM Institution, have a laboratory account, synchronize the data acquired on the platform daily and share them to their users.

The detailed workflow offered to NEP laboratories is available at [15], along with the Nextcloud desktop client installation guide [16]. Moreover, users can take advantage of the video tutorial [17], and the FAQs [18].

Electronic lab notebook (eLabFTW)

NFFA Europe Pilot infrastructure does not include its own Electronic Laboratory Notebook (ELN), but the project's Research Data Policy recommends the use of it if the Institution provides one, as stated at point 5.1.7 of NFFA-Europe Research Data Policy [12].

The use of an ELN by the Laboratories within the infrastructure is monitored through the Lab-DMPs. At present, about half of the laboratories that provided a DMP are using an ELN, of which about 2/3 are adopting eLabFTW.

eLabFTW is an open-source ELN, developed since 2012, and distributed under AGPL-3.0 license. It is written in the PHP language and supports the RESTful APIs. In addition, a Python library, named *elabapy*, provides APIs to work with tools developed in Python language.

eLabFTW allows users to document their research activity through the so-called "Experiments" which are described in a simple text field and can be further enhanced with pictures and drawings (HTML or MD).

It also provides an inventory, called "Database", to collect the Items used in the laboratory, which can be differentiated into categories (e.g. instruments, equipment, substrates, chemicals, but also as general as the publications of the group). All these Items can be linked between them and with the Experiments and both Experiments and Items support the addition of attached metadata to it in JSON format.

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

Personal data management

As stated in NEP Data Management Plan [19], authentication to the data management tools used in the project (NFFA Datashare, Data Stewardship Wizard and MetaRepo) are managed by the single sign-on system and stored exclusively in the Identity Provider database (Keycloak), while only an anonymous identifier is propagated. The MetaRepo collects monitoring data (ID of the authenticated user and the ID of the service) in order to send them to the NEP backend, without storing them locally.

A set of non-sensitive personal data of the members of the proposing research team (user ID, email, first name, last name, affiliation, country, role of the research user in the proposal) are mapped automatically by the NEP central database and registered in the MetaRepo after an accepted proposal is assigned to the access provider(s), and some limited personal information can optionally be provided by the author in the metadata document describing the datasets generated within NEP uploaded on the MetaRepo. The purpose of this personal data is uniquely acknowledging the authorship of the data and related metadata for proper citation of the research work.

The handling of personal data is regulated by the Privacy Policy (https://www.nffa.eu/apply/privacy) and the Data Sharing Agreement accepted by all Beneficiaries and Third Parties, which governs the Joint Controllerships. All personal data are retained with the scope of delivering the service and the project follows the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 – general data protection regulation (GDPR).

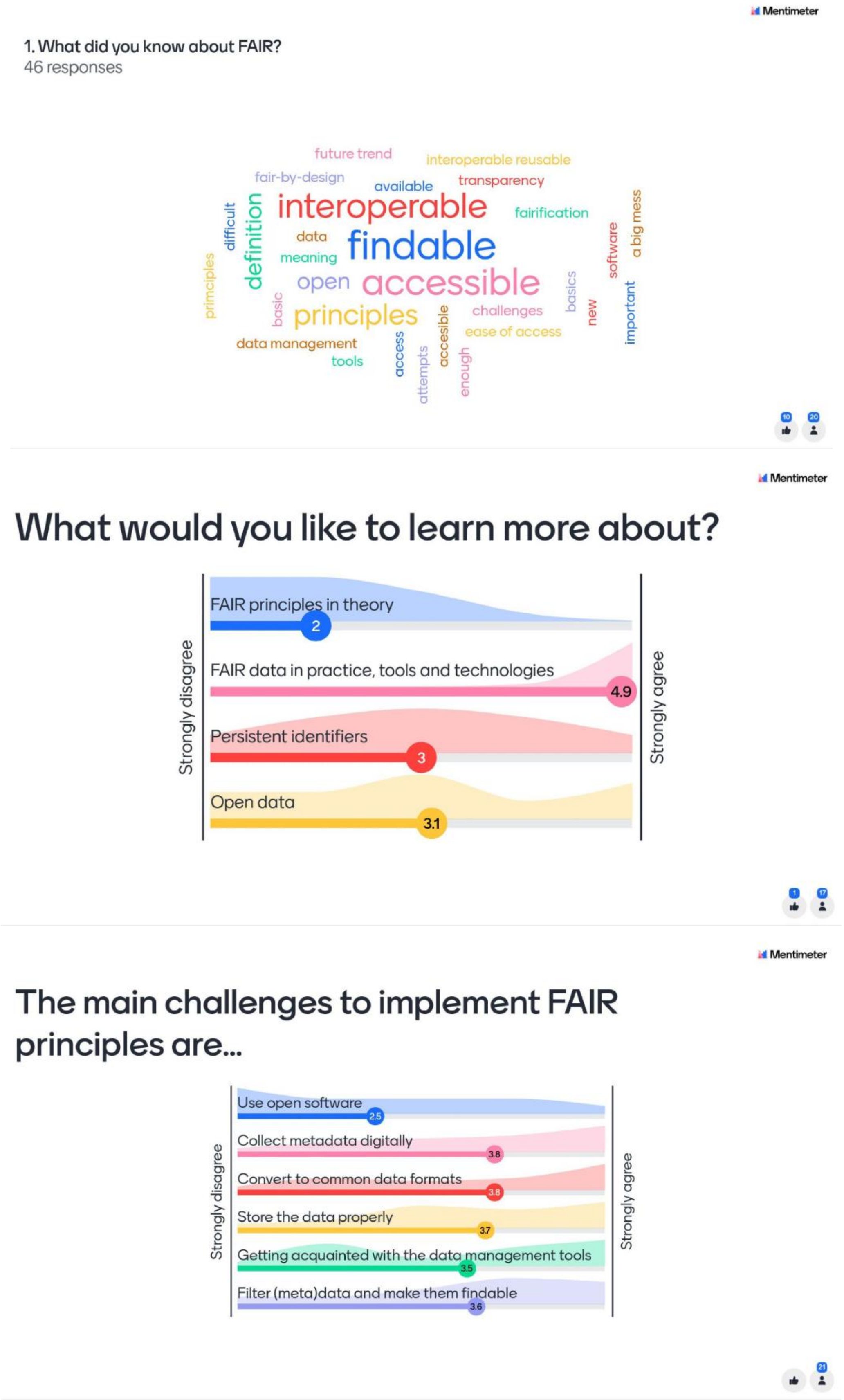
Data management dissemination and users' feedback

The approach and tools devised in NFFA Europe Pilot to address FAIR data management are currently being disseminated in various ways. First, the public workshop "FAIR and Open Data in NFFA-Europe Pilot and Beyond" [20] was held on September 26th 2023 in Garching bei Munchen, as a satellite event of the project General Assembly. The workshop presented the activity carried on in NFFA Europe Pilot about data management, and included four invited talks from other communities and projects involved in data management, namely FAIRmat, NFDI-MatWerk, NCCR Marvel, MAX CoE and GO NANO/GO FAIR. The workshop had more than 20 participants in person and a peak of above 50 participants remotely.

Then, a comprehensive set of video tutorials will be developed to guide users through each tool's functionality, covering the entire spectrum from laboratory procedures to search methodologies. Concise 10-second announcement videos will be created to generate interest and awareness. The tutorials will be structured with varying levels of complexity, catering to beginners, intermediate learners, and advanced users. To appeal to a wider audience, we will incorporate multiple use cases. This will be accomplished through a combination of screen recordings and recordings in laboratories, and covering various laboratories that deal with different types of data. The scheduled release of these videos in spring 2024 aims to expedite knowledge dissemination. Furthermore, the presentations [21] and recordings [22] from "FAIR and Open Data in NFFA-Europe Pilot and Beyond" workshop and the recordings for future schools and workshops will be shared on NFFA YouTube channel, providing an extended platform for knowledge sharing and engagement.

Finally, a three-day school will be organized tentatively in the first half of June 2024. The program will cover the basics of data management theory and practical applications, focusing on tools for data management and virtual access services, offering both tutorials and hands-on experiences. With a target audience of 60 participants, the program will be shaped by feedback from "FAIR and Open Data in NFFA-Europe Pilot and Beyond" workshop, ensuring relevance and addressing specific needs identified in the field.

The workshop gave us the chance to get feedback from the audience about their feelings, expectations and problems with the application of FAIR data management. The feedback has been collected using a live pool by means of a set of questions on MentiMeter. In fig 3 we report the questions and the statistics of the answers.


Mentimeter

1. What did you know about FAIR?
46 responses

Mentimeter

What would you like to learn more about?

Mentimeter

The main challenges to implement FAIR principles are...

Strongly disagree
Strongly agree
Use open software
2.5
Collect metadata digitally
3.8
Convert to common data formats
3.8
Store the data properly
3.7
Getting acquainted with the data management tools
3.5
Filter (meta)data and make them findable
3.6
21

Figure 3: feedbacks from “FAIR and Open Data in NFFA-Europe Pilot and Beyond” workshop

From the results, it emerges that the participants are familiar with the core principles of FAIR. Some acknowledge the difficulties and complexities associated with implementing FAIR, as evidenced by

This initiative has received funding from the EU’s H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

phrases like "difficult," "fair-by-design," and "a big mess." There's also an awareness of the evolving aspects of FAIR adoption, as indicated by mentions of "transparency," "software," and "future trend." A notable emphasis on exploring the practical facets of FAIR, such as its application, tools, and technologies, is evident. This underscores a practical orientation and a keen interest in gaining hands-on knowledge. Furthermore, there is interest in understanding persistent identifiers and open data, indicating a comprehensive curiosity about various aspects of FAIR principles.

Participants stress the importance of digitally collecting metadata, converting data into common formats, and ensuring proper data storage. These challenges highlight an acknowledgment of the technical complexities involved in adopting FAIR. Additionally, the need to filter (meta)data for findability reflects a practical concern for making FAIR principles actionable.

References

[1] NFFA Datashare: https://datashare.nffa.eu

[2] Metadata editor download and documentation: https://metadata-editor.gitlab.io/documentation/

[3] MetaRepo homepage: https://metarepo.nffa.eu

[4] NEP Data Stewardship Wizard instance https://dsw.nffa.eu/

[5] Könnecke, M., Akeroyd, F.A., Bernstein, H.J., Brewster, A.S., Campbell, S.I., Clausen, B., Cottrell, S., Hoffmann, J.U., Jemian, P.R., Männicke, D., Osborn, R., Peterson, P.F., Richter, T., Suzuki, J., Watts, B., Wintersberger, E., Wuttke, J.: The NeXus data format. Journal of Applied Crystallography 48(1), 301–305 (2015)

[6] MetaStore documentation: https://github.com/kit-data-manager/metastore2/blob/metastoreGui/restDocu.md

[7] Aversa, R., Brandino, G. P., Chelbi, S., Hartmann, V., Jejkal, T., NFFA-Europe Pilot - D9.1 - Deployment of the VA service prototypes. DOI: 10.5281/zenodo.7702265

[8] Aversa, R., Panighel, M., Recchia, A., Chiucconi, Y., Brandino, G. P., Tsibidis, G., Pantazis, Y., Sfakianakis, I., Grassano, D., NFFA-Europe Pilot - D9.2 - Integration of the first VA services into the platform. DOI: 10.5281/zenodo.10156330

[9] Blumenröhr, N., MacKinnon, N., Aversa, R., FAIR Data Management Workflow for MRI Data, Proceedings of the 3rd Workshop on Metadata and Research (objects) Management for Linked Open Science (DaMaLOS 2023). DOI: 10.4126/FRL01-006444992

[10] Blumenröhr, N., Aversa, R., From implementation to application: FAIR digital objects for training data composition. Research Ideas and Outcomes 9: e108706. DOI: 10.3897/rio.9.e108706

[11] Aversa, R., Blumenröhr N., Recchia, A., Tsibidis, G.D., Brandino, G. D., Pantazis, Y., NFFA-Europe Pilot - D16.4 - Report on additional data services. DOI: 10.5281/zenodo.10201227

[12] NFFA-Europe Research Data Policy, https://www.nffa.eu/apply/data-policy

[13] NEP page on OpenAIRE: https://explore.openaire.eu/search/project?projectId=corda__h2020::eaec8ea099430906ba4e6df66a986a03

[14] Zenodo community “NFFA-Europe Pilot project material”: https://zenodo.org/communities/nffa-europe-pilot

[15] NFFA Datashare workflow for Laboratories: https://datashare.nffa.eu/s/NwarkwLQQd7CoHr

[16] NFFA Datashare installation guide for Laboratories: https://datashare.nffa.eu/s/CB8dBkm8tmdn35c

[17] NFFA Datashare video tutorial for users: https://www.youtube.com/watch?v=qSlq6Jesn5M

[18] NFFA Datashare FAQ: https://datashare.nffa.eu/index.php/s/Mo97YsxXYDSJMGZ

[19] Modolo, I., Osmenaj, E., Cozzini, S., Piacini, E., Panighel, M., Narducci, E., Brandino,G.P., Aversa, R., NFFA-Europe Pilot - D1.1 - Data Management Plan (DMP). DOI: 10.5281/zenodo.7701973

[20] "FAIR and Open Data in NFFA-Europe Pilot and Beyond" event: https://www.nffa.eu/news/events/two-joint-nffa-europe-workshops-on-microscopy-open-data-and-analytical-large-scale-facilities/

[21] Presentations of the workshop "FAIR and Open Data in NFFA-Europe Pilot and Beyond" https://zenodo.org/communities/nffa-europe-workshop-fair-2023/

[22] Recording of the workshop "FAIR and Open Data in NFFA-Europe Pilot and Beyond" https://datashare.nffa.eu/s/T8DSQkKQFrF7bYs