nffa.eu

PILOT 2021 2026

DELIVERABLE REPORT

WP16 JA6 - Implementing FAIR data approach within NEP

D16.4

Report on additional data services

Due date

M30

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
WP16	JA6 – Implementing FAIR data approach within NEP

WORK PACKAGE LEADER

Giuseppe Piero Brandino (eXact lab)

DELIVERABLE DETAILS

DELIVERABLE ID	DELIVERABLE TITLE
D16.4	Report on additional data services

DELIVERABLE DESCRIPTION

This document describes the new data services developed within NFFA Europe Pilot, extending the list of data services provided in Deliverable D16.2 at Month 18. Some of these services will be provided as Virtual Access at Month 31 or at Month 37, according to the schedule agreed in the proposal.

DUE DATE	ACTUAL SUBMISSION DATE
M30 31/08/2023	04/10/2023

AUTHORS

Rossella Aversa (KIT), Nicolas Blumenröhr (KIT), Andrea Recchia (eXact lab), G.D. Tsibidis (FORTH), Y. Pantazis (FORTH)

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

PERSON RESPONSIBLE FOR THE DELIVERABLE

Giuseppe Piero Brandino (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
04/10/2023	**14** (right-click and select "update the field")

FOR MORE INFO PLEASE CONTACT

Giuseppe Piero Brandino
eXact lab
via Crispi, 56 – 34126 Trieste
(Italy)

email: brandino@exact-lab.it

VERSION	DATE	AUTHOR(S)	DESCRIPTION / REASON FOR MODIFICATION	STATUS
1	31/07/2023	R. Aversa, N. Blumenröhr	First draft	Draft
2	04/08/2023	A. Recchia	Section 5	Draft
3	10/08/2023	R. Aversa, N. Blumenröhr	Section 1	Draft
4	11/08/2023	R. Aversa, N. Blumenröhr	Section 2	Draft
5	30/08/2023	G. D. Tsibidis, R. Aversa	Section 4	Draft
6	31/08/2023	Y. Pantazis, R. Aversa	Section 3	Draft

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

INTRODUCTION TO DATA SERVICES

This deliverable presents a description of the new data services developed by the Work Package 16 within NFFA Europe Pilot, extending the list of data services provided in Deliverable D16.2 at Month 18. All the services are intended to improve the FAIRness of the data, either by post-processing of already existing data or by design at new data generation. Some of these services are planned to be included in the Virtual Access offer at Month 31 or at Month 37, according to the schedule agreed in the Proposal. The document is structured in sections describing one data service each. For completeness, each section explicitly includes the details about the future Virtual Access developments of the service, if this has been planned.

1. Magnetic Resonance image reconstruction and contrast prediction

Magnetic Resonance Imaging (MRI) is applied in material sciences for non-invasive investigation of sample structure and composition, by leveraging the differences in tissue contrasts. However, every different type of contrast, encoded in the MR image, typically requires a separate measurement, which is a time-consuming task. In fact, the contrast in an MRI image emerges from the differences between pixel intensities $I1(x_1,y_2)$ and $I2(x_1,y_2)$ which depend on the signal that is specific for the sample properties and the parameters of the MR pulse sequence. The relevant pulse sequence parameters are the echo time (TE) and the repetition time (TR). They are correlated with the contrast types, i.e., T1-, T2-, and proton density- (PD) weighted contrast. The image contrast is weighted in dependence of the sample properties and the length of the sequence parameters relative to the length of the physical contrast values. Since the true T1, T2 and PD values are typically unknown for a given sample, are affected by the measurement device, and can vary between different sample tissues, enhancing the contrast between various areas in the image requires multiple measurements with alternating TE and TR values, which are part of a specific MRI sequence protocol.

The aim of this service is to optimise the information contained within the datasets measured in the same MRI experiment, to predict an alternative contrast type from a given one. This decreases the experimental time by a factor n for each of the n contrast types that can be predicted.

To reduce the number of measurements, the proposed solution is to employ Machine Learning (ML) techniques for image contrast prediction. The essential idea is to train a ML model capable of finding relations between the contrast weights of the same image. Given a measured contrast, the model can then be used to predict, i.e., generate, the image for an alternative contrast.

Technically, this problem can also be addressed as an image transformation task. The most common techniques applied in this field are Convolutional Neural Networks (CNN), which are especially suited for image data. The U-net structure, originally used in the field of medical image segmentation [1], has been proved as particularly efficient and is a promising candidate for this challenge.

The T1-, T2-, and PD- weighted image contrasts are fluid, and some images will often also contain contributions from the other contrasts when dominated by a specific one. Thus, a better approach is to express the contrast type in terms of the pulse sequence parameters TE and TR applied to a known sample type. A U-net will therefore receive the measured image together with the numerical values of the pulse sequence parameters for the given image, in order to predict the image for the alternative pulse sequence parameters. If trained with sufficient data points, the model should be capable of predicting multiple contrasts when pairs of images with alternating contrast types were used. Possible combinations are T1- to T2-w., T1- to PD-w., T2- to PD-w., and vice versa.

However, the system becomes rapidly more complex with an increasing number of sample types, sequence protocols, and measurement devices used to generate the images. Without providing additional features that are related to these parameters, the model must learn the effects on the differences between multiple contrasts of the images. Therefore, the best practice is to start with a simple system where only one sample type, one sequence protocol, and one measurement device has been used to generate images with different contrast types. The model complexity can be gradually increased by adding data points of images with different provenance and sample characteristics. Then, the model performance is compared to the prior state.

Currently, a preliminary U-net model [2] has been trained with a dataset of 138 series, with each series containing 1 or 10 images per sequence protocol and having a different set of TE and TR values. These images were recorded at the institute of Microstructure Technology at the Karlsruhe Institute of Technology and show a set of 7 sample tubes that contain different concentrations of $CuSO_4$ (i.e., 0, 5, 10, 25, 50, 75, and 100 millimolar). Part of the used dataset is openly available [3] while the remaining data is still under embargo. Additionally, the number of images was artificially increased by image augmentation methods, varying the sample positions. In order to predict the alternative image contrast from a given one, images measured with specific TE and TR values were grouped into T1- and T2-weighted image contrast groups and aligned pairwise for training. Furthermore, the corresponding TE and TR values of the input and output image were used as additional features for the U-net.

In its current version, the user may input in the model [2] the measured image, the pulse sequence parameters of the measured image, and those of the expected target image with the new contrast weighting. The output is the predicted image with the new contrast.

As an example, Figure 1 shows the preliminary result obtained using as input a MRI image measured with TE=8ms, TR=200 ms (corresponding to a T1-weighted estimated contrast) and TE=25 ms, TR 500=ms (corresponding to a T2-weighted estimated contrast) as the parameters of the target image to be predicted. The comparison of Figure 1b and Figure 1c gives a qualitative impression about the success of the prediction algorithm.

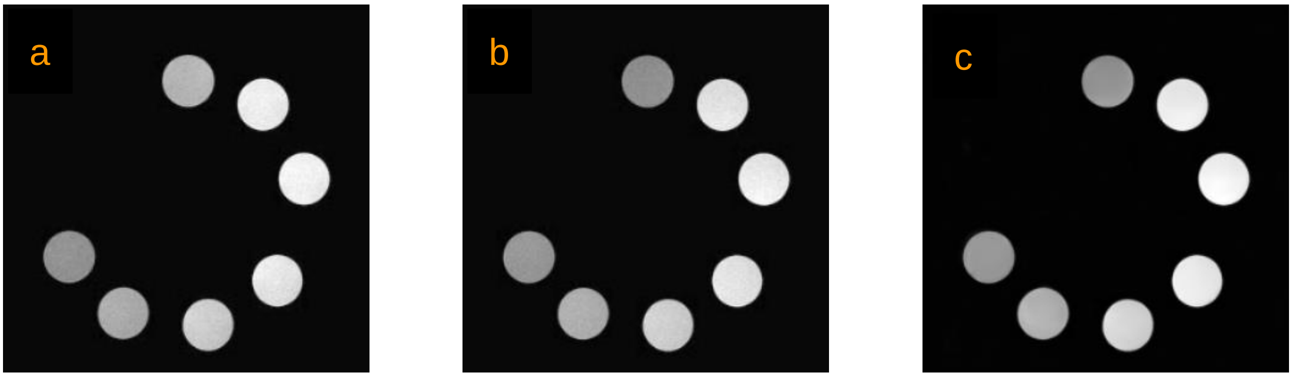


Figure 1: Preliminary results obtained with the contrast prediction model. Input image (a), ground truth (b), predicted image (c)

To further improve the model, it is planned to increase the training data for the same or even additional samples, measured with specific TE and TR values that are currently missing. Furthermore, the U-net architecture will also be optimised. The trained model will be offered as a Virtual Access service for authenticated users starting from Month 37.

2. Nuclear Magnetic Resonance data curation

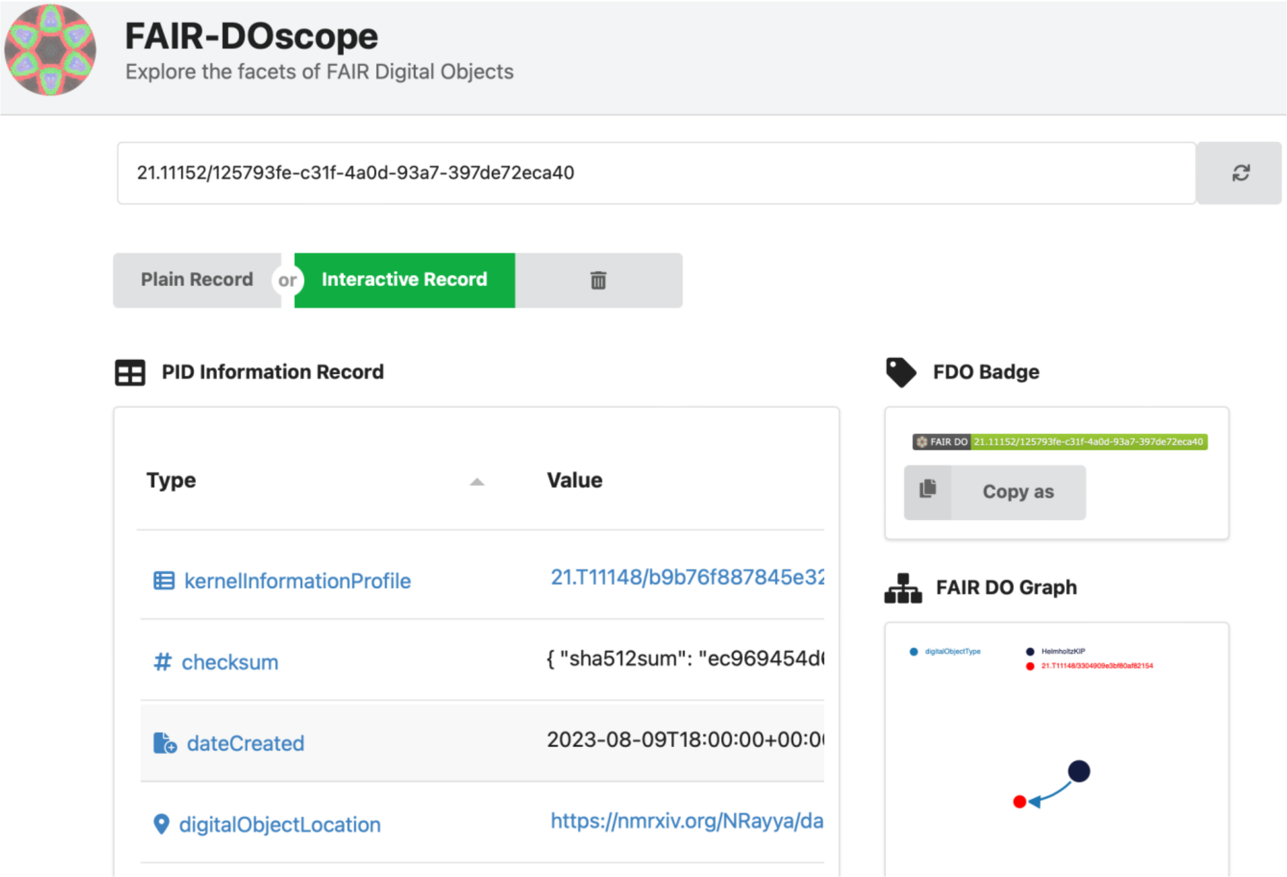
Data curation is the first step towards improving Nuclear Magnetic Resonance (NMR) correlation and spectral analysis algorithms, as well as enabling automation. NMR databases (e.g., [4]) and structured data formats (e.g., [5]) have been explored by many initiatives. The challenge, however, is to identify a data representation that not only contains all the essential data to reliably reconstruct the original spectra and to re-calculate the resonances, but also allows to uniformly compare and reuse data archived in different locations.

We propose to use the concept of FAIR Digital Object (FDO) for the NMR spectra, as well as for the corresponding metadata. A FDO [6] is a data representation identified by a globally-unique, resolvable, Persistent Identifier (PID), described by an information record, and classified by a type which determines the operations a machine can perform on the data (e.g., in this case, the reconstruction algorithm, specified as FDO itself). The information record, created following the Helmholtz Kernel Information Profile (KIP) [7], contains the information required to access the data, in particular the reference to the repository where the data is deposited.

This approach enables repository-agnostic (meta)data interpretation on a common basis for machines, connecting and relating the original data without any direct changes or any data migration. Thus, irrespective of where the NMR spectra are located at, they can be identified, accessed, and reused. With the introduction of this additional abstraction layer, the required operations for reconstructing the spectra and re-calculating the resonances can be included in the description (information record) of the FDO representing the original data in a structured and standardised way.

FDOs are primarily intended for machine-actionability and automation; nevertheless, a human-readable format of the information record is also possible. For this reason, the FAIR-DOscope [8] has been developed, which is a generic FDO viewer and browser which offers a tabular view of the contents of the information record and a graphical representation of related FDOs.

As an example, we created the FDO for the Caripyrin - NMR Spectra Dataset [9] deposited in nmrXiv [10] and for its associated publication [11]. The PIDs [9, 11] can be used as input in the FAIR-DOscope [8] to visualise their content. Figure 2 shows how the FDO for the Caripyrin - NMR Spectra Dataset is intuitively represented in FAIR-DOscope.



FAIR-DOscope

Explore the facets of FAIR Digital Objects

21.11152/125793fe-c31f-4a0d-93a7-397de72eca40

Plain Record
or
Interactive Record

PID Information Record

Type	Value
kernelInformationProfile	21.T11148/b9b76f887845e32
checksum	{ "sha512sum": "ec969454d(
dateCreated	2023-08-09T18:00:00+00:0(
digitalObjectLocation	https://nmrxiv.org/NRayya/da

FDO Badge

FAIR DO 21.11152/125793fe-c31f-4a0d-93a7-397de72eca40
Copy as

FAIR DO Graph

digitalObjectType
HelmholtzKIP
21.T11148/3304909e3bf80af82154

Figure 2: A partial screenshot highlighting the tabular view of the information record in human-readable form (left) and the graphical representation of the related FDOs (right) for the Caripyrin - NMR Spectra Dataset in FAIR-DOscope

As a future development, an interface to a FDO graph will be provided, which can be used to list the collection of available FDOs representing the NMR spectra and the corresponding information. Users will be able to formulate SPARQL queries to search and filter the FDO graph in order to retrieve the data and metadata according to use-case specific criteria. The tool will be offered as a Virtual Access service for authenticated users starting from Month 37.

3. Predictive service for nanostructures

The presented service allows to predict morphological nanostructure of laser-processed surfaces from Machine Learning (ML) models, which have been trained on annotated datasets. The use of ML predictive models is motivated by the fact that they are computationally efficient, and their interpolation accuracy is very high.

The predictive model of the service takes an input the natural logarithm of the number of pulses and the fluence, and returns the most probable surface structure. Feature engineering is performed, and new higher-order features are constructed and then fed to the ML models. The software trains several ML models including k-Nearest Neighbours, Gaussian Naive Bayes, Logistic Regression, Support Vector Classifier and Gradient Boosting Classifier. Due to the small size of the datasets, neural networks and other data-hungry predictive models were excluded. The model has been implemented in Python using the scikit-learn library [12].

There are two usage scenarios for the predictive service:

1. The use of trained ML models on new input data to predict their most probable surface pattern. The available predictive models have been trained on the following materials: Si, Steel 1.7131, Steel 1.7225 and Ti6Al4V. The expected input file (in csv format) should contain two columns corresponding to the input variables (natural logarithm of the number of pulses and fluence).
2. The training of the ML models on new materials. The expected input file (in csv format) should contain three columns: natural logarithm of the number of pulses, fluence, and the label of the ground-truth surface pattern corresponding to the input variables. All the available ML models are then trained on the new input data and a final report, containing the mean accuracy and the corresponding standard deviation from a 5-fold cross-validation procedure, is generated. The best model is saved and can be used at later times for predictions (usage scenario 1) on the same material.

A Graphical User Interface is currently under construction, and it will be available for authenticated users as part of the Virtual Access offer starting from Month 31.

4. Materials Modelling: Damage Threshold Evaluation

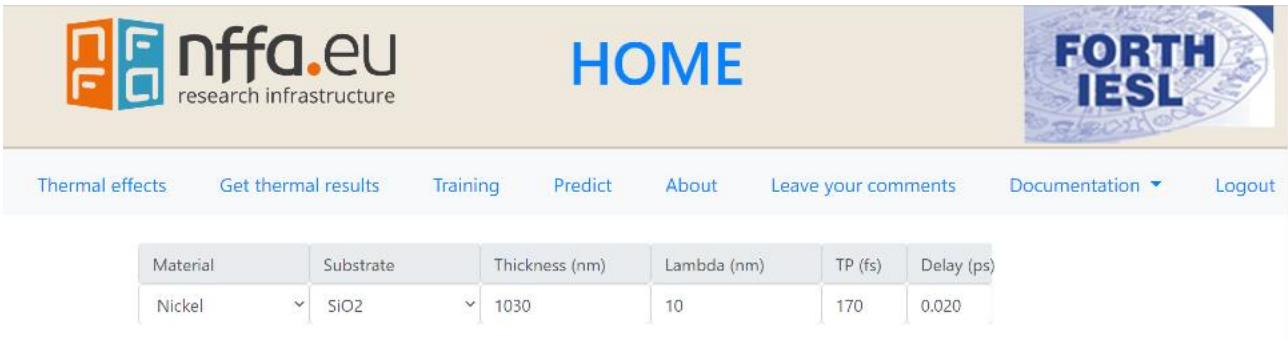
The employment of femtosecond (fs) pulsed lasers has received significant attention due to its capability to facilitate fabrication of precise patterns at the micro- and nano- lengths scales. A key issue for efficient material processing is the accurate determination of the damage threshold that is associated with the laser peak fluence at which minimal damage occurs on the surface of the irradiated solid. Despite a wealth of previous reports that focused on the evaluation of the laser conditions that lead to the onset of damage, the investigation of both the optical and thermal response of thin films of sizes comparable to the optical penetration depth is still an unexplored area.

To describe the damage induced on the material following irradiation with fs pulses, a theoretical framework is employed to explore the excitation and thermal response of a double-layered structure (thin metal film on a dielectric material). The simulation algorithm is based on the use of a Two Temperature Model (TTM) that represents the standard approach to evaluate the dynamics of electron excitation and relaxation processes in solids [13]. For the sake of simplicity, a 1D-TTM is used to describe the thermal effects due to heating of the thin films with laser pulses of wavelength λ_L for a pulse duration τ_P. This multiscale physical model is used to correlate the energy absorption, electron excitation, relaxation processes and minimal surface modification [14, 15]

A Graphical User Interface (Figure 3) has been developed to allow a user-friendly evaluation of the impact of various parameters such as the photon energies, the pulse duration, the pulse separation (in case of double pulse experiments) and the material thickness on the damage

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

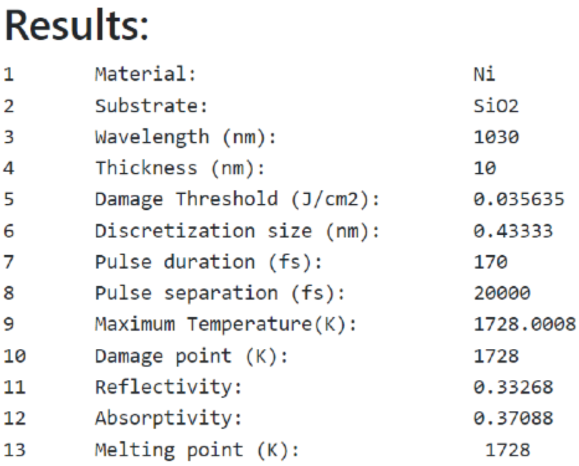
threshold for various metals (Au, Ag, Cu, Al, Ni, Ti, Cr, Stainless Steel); three different substrates (Si, SiO2 and soda lime silica glass) have been considered.



Material	Substrate	Thickness (nm)	Lambda (nm)	TP (fs)	Delay (ps)
Nickel	SiO2	1030	10	170	0.020

Figure 3: Interface of the Damage Threshold Evaluation

The software that has been developed to perform the simulations is written in Octave and the execution of the code is conducted on fast computers located at FORTH. Results are saved in txt format files (Figure 4) that contain the prescribed laser conditions and the materials used as well as information of the damage threshold (i.e. minimum fluence to melt the material). The maximum temperature in these conditions and comparison with the melting temperature is used to assess the accuracy of the algorithm. Furthermore, optical parameters (i.e. reflectivity and absorptivity) of the irradiated metals are calculated and their values are included in the txt file. A documentation [16] has also been included in the interface to provide the steps a user should follow as well as the underlying theory that describes the multiscale physical processes that lead to material damage.



Results:

```
Material:                        Ni
Substrate:                       SiO2
Wavelength (nm):                 1030
Thickness (nm):                  10
Damage Threshold (J/cm2):        0.035635
Discretization size (nm):        0.43333
Pulse duration (fs):             170
Pulse separation (fs):           20000
Maximum Temperature(K):          1728.0008
Damage point (K):                1728
Reflectivity:                    0.33268
Absorptivity:                    0.37088
Melting point (K):                1728
```

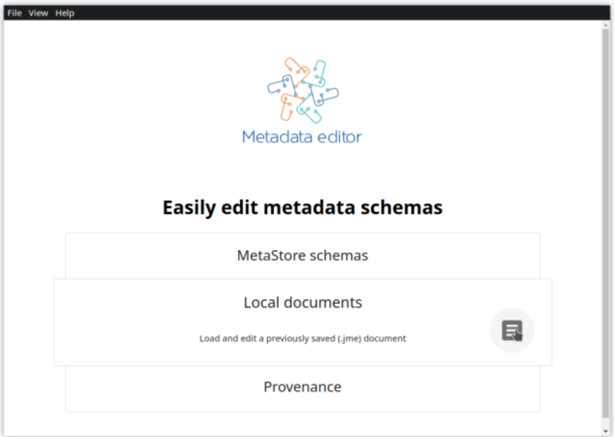
Figure 4: Output of the simulation for the Damage Threshold Evaluation

The execution of the algorithm takes into account both the thermophysical and optical parameter values of the irradiated complex (metal/substrate). It has been developed assuming the most common metals and substrates while it is aimed to be extended soon to other metals and configurations (multi-layered materials, etc). The tool will be offered as a Virtual Access service for authenticated users starting from Month 31.

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

5. Metadata Editor

The Metadata Editor is an open-source, cross-platform desktop application designed to efficiently retrieve metadata schemas from MetaStore and compile metadata documents [17]. This tool gives to the users an intuitive interface (Figure 5) to register metadata documents to MetaStore or save/export them for future modifications and reuse. To facilitate the automation and the interaction with experimental instruments and electronic notebooks, the application provides REST endpoints for schema compilation and field reading.



Easily edit metadata schemas

Load and edit a previously saved (.jme) document

Figure 5: Metadata Editor home page

The main workflows for the Metadata Editor, illustrated in Figure 6, are:

1. Generate a metadata document:

    1.1. Choose one of the following options:

        1.1.1. Select a metadata schema from the MetaStore. It is possible to filter them by label, ID, and version.

        1.1.2. Load a previously saved .jme file to complete or modify the metadata document compilation

    1.2. Compile the rendered form. During this stage, the tool automatically validates the input and displays any errors or missing mandatory fields to the user.

2. Save a metadata document:

        2.1.1. Upload the metadata document to MetaStore. In this case, the user needs to provide a metadata schema record with relevant information about the compiled metadata document. NFFA login is required for this step.

        2.1.2. Export the generated document as a .json file.

        2.1.3. Save a file in a proprietary format (.jme) for completing the schema compilation later.

3. Obtain the provenance:

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

3.1. Select the metadata document for which to generate the provenance. The documents are grouped by schema ID and can be chosen from a dropdown menu, sorted by last update date and time. NFFA login is required for this step.

3.2. Save the provenance for the selected document as a .json file.

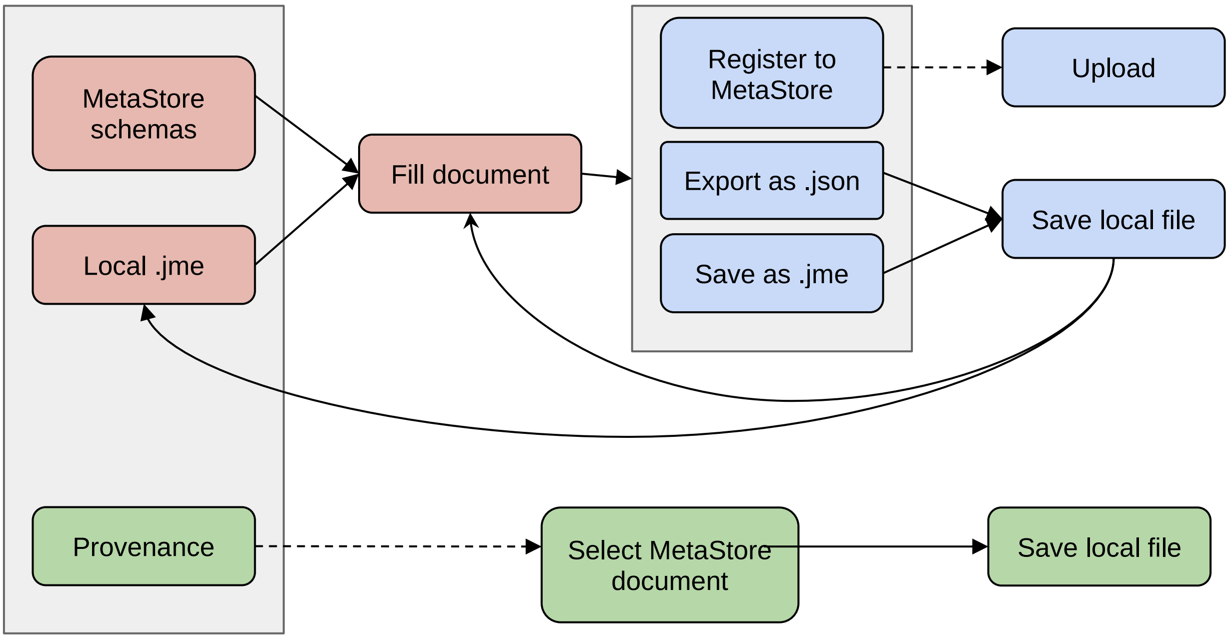
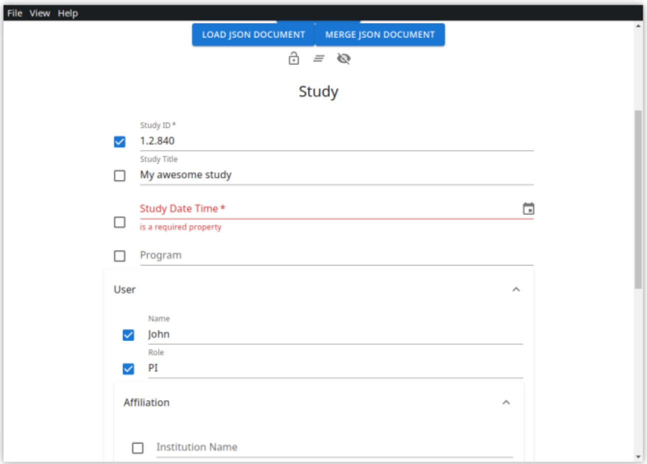


Figure 6: Metadata Editor workflow. Red, blue and green processes indicate flow 1, 2 and 3 described in the text, respectively. Dashed arrows indicate that authentication is required. Process inside grey boxes are on the same application page and are mutually exclusive

The editor offers a set of useful tools to enhance the form visualisation and compilation process, as depicted in Figure 7. The user can selectively 'lock' the fields by checking them individually. This allows for hiding the unlocked fields, leaving only the locked ones visible. This is particularly handy because certain fields could be automatically populated from a locally pre-filled JSON or from data provided by some experimental instrument, and this locking functionality enables immediate access to only the relevant fields for editing without the need to scroll through the entire form.



Study

Study ID *
☑ 1.2.840

Study Title
□ My awesome study

□ Study Date Time *
is a required property

□ Program

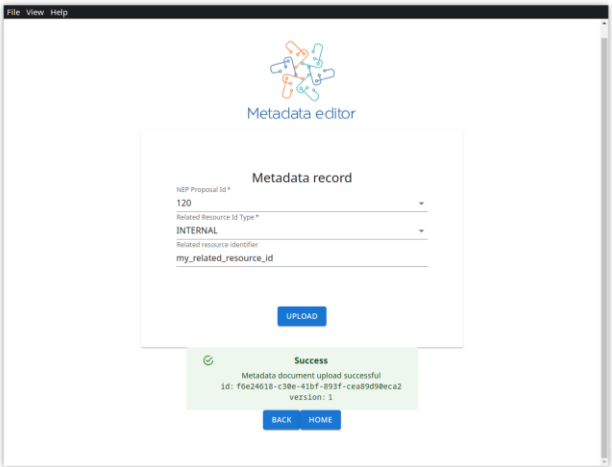
User

Name
☑ John

Role
☑ PI

Affiliation

□ Institution Name



Metadata record

NEP Proposal Id *
120

Related Resource Id Type *
INTERNAL

Related resource identifier
my_related_resource_id

Success
Metadata document upload successful
id: f6e24618-c30e-41bf-893f-cea89d90eca2
version: 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 7: Metadata Editor screenshots for some steps of flow 1. Form compilation (left) and metadata document uploading (right)

Additionally, local .json documents can be loaded to populate the form by using the 'load .json document' button. Furthermore, these documents can be utilised to create a document merge, where only the locked fields remain untouched, and all other fields are overwritten. This can be done through the 'merge .json document' button. This functionality is particularly useful when multiple partially filled .json documents already exist and need to be merged, e.g. the "user" information and the "instrument" parameters.

Technical details

The metadata editor has been developed using the Electron framework, starting from the Electron React Boilerplate project [18], with React.js and the JSONForms library to generate compilable forms from JSON schemas [19]. For defining and exposing REST endpoints, Express.js is used.

To generate provenance and register metadata documents to MetaStore, users must authenticate using NFFA credentials. The login procedure is managed by utilising the AppAuth-JS library, which ensures a secure flow of authorization by employing the user's browser and PKCE protocol [20]. This approach adheres to the RFC 8252 best practices for OAuth 2.0 in Native Apps [21]. No personal data is stored or handled by the Metadata Editor, the keycloak uuid and token being the only information needed to grant access.

Being a desktop application, the tool is currently not part of the Virtual Access, and there are no plans to include it in the future.

References

[1] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, Vol.9351: 234—241 (2015). DOI: 10.48550/arXiv:1505.04597

[2] N. Blumenröhr, MRI contrast prediction package https://github.com/nicolasblumenroehr/MRI-Contrast-Prediction-Package

[3] N. Blumenröhr, N. MacKinnon, R. Aversa, Magnetic Resonance Imaging Copper Sulfate Dataset. Zenodo (2022). DOI: 10.5281/zenodo.7319761

[4] nmrXiv webpage https://nmrxiv.org/

[5] NMReDATA Initiative webpage https://www.nmredata.org/

[6] E. Schultes, & P. Wittenburg, FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. In Y. Manolopoulos & S. Stupnikov (Eds.), *Data Analytics and Management in Data Intensive Domains* (Vol. 1003, pp. 3–16). Springer International Publishing (2019). DOI: 10.1007/978-3-030-23584-0_1

[7] Helmholtz Metadata Collaboration, Helmholtz Kernel Information Profile. HMC Paper, 2. 35 pp. DOI: 10.3289/HMC_publ_03

[8] FAIR-DOscope access page https://kit-data-manager.github.io/fairdoscope/

[9] Caripyrin - NMR Spectra Dataset. PID: 21.11152/125793fe-c31f-4a0d-93a7-397de72eca40

[10] Caripyrin - NMR Spectra Dataset. DOI: 10.57992/nmrxiv.p7

[11] P.H. Rieger, J.C. Liermann, T. Opatz, H. Anke, E. Thines, Caripyrin, a new inhibitor of infection-related morphogenesis in the rice blast fungus Magnaporthe oryzae. PID: 21.11152/f1291732-9d19-4284-b35a-25b6804d4ff3

[12] https://scikit-learn.org/stable/

[13] S. I. Anisimov, B. L. Kapeliovich, and T. L. Perel'man, Zhurnal Eksperimentalnoi Teor. Fiz. 66, 776 (1974) [Sov. Phys. Tech. Phys. 11, 945 (1967)]

[14] G.D. Tsibidis, E. Stratakis, 'The impact of the substrate on the opto-thermal response of thin metallic targets following irradiation with femtosecond laser pulses', Journal of Central South University 29, 3410 (2022). DOI: 10.1007/s11771-022-5169-4

[15] G.D. Tsibidis, E. Mansour, E. Stratakis, 'Damage threshold evaluation of thin metallic films exposed to femtosecond laser pulses: the role of material thickness', Optics and Laser Technology 156,108484 (2022). DOI: 10.1016/j.optlastec.2022.108484

[16] G. Tsibidis, Manual for the use of the interface to compute the damage threshold of metals of various thicknesses following irradiation with femtosecond laser pulses. https://nffa-modeling.iesl.forth.gr/docthermal

[17] Metadata Editor GitLab page https://metadata-editor.gitlab.io/documentation/

[18] https://electron-react-boilerplate.js.org/

[19] https://jsonforms.io/

[20] https://github.com/openid/AppAuth-JS

[21] https://datatracker.ietf.org/doc/html/rfc8252