DELIVERABLE REPORT

WP11 NA – Innovation and networking activities

D11.14

Final metadata standard for nanoscience data

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
WP11	NA – Innovation and networking activities

WORK PACKAGE LEADER

Ennio Capria (ESRF)

DELIVERABLE DETAILS

DELIVERABLE ID	DELIVERABLE TITLE
D11.14	Final metadata standard for nanoscience data

DELIVERABLE DESCRIPTION

This document contains the NFFA Deliverable D11.14 “Final metadata standard for nanoscience data” due in M30.

EXPECTED DATE	ESTIMATED INDICATIVE PERSONMONTHS
M30 28/02/2018	8 MM

AUTHOR(S)

Vasily Bunakov (vasily.bunakov@stfc.ac.uk), Brian Matthews (brian.matthews@stfc.ac.uk), Thomas Jejkal (thomas.jejkal@kit.edu), Rossella Aversa (aversa@iom.cnr.it), Stefano Cozzini (cozzini@iom.cnr.it)

PERSON RESPONSIBLE FOR THE DELIVERABLE

Ennio Capria (ESRF)

NATURE

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

REPORT DETAILS

ACTUAL SUBMISSION DATE

28/02/2018

NUMBER OF PAGES

42(right-click and select “update the field”)

FOR MORE INFO PLEASE CONTACT

Brian Matthews (STFC) Tel. +44 01235 446648 Email: brian.matthews@stfc.ac.uk

Version	Date	Author(s)	Description / Reason for modification	Status
1	11/01/2018		Initial compilation of work	Draft
2	24/01/2018		Implementation section added	Update
3	7/02/2018		Appendixes added	Update
4	19/02/2018		Reviewed internally	Update
5	26/02/2018		References corrected	Final
6	28/02/2018	Brian Matthews	Tables ad figures’ references corrected	Final

Contents

Executive summary

This document contains the NFFA Deliverable D11.14 “Final metadata standard for nanoscience data” due in M30. It contains references to the NFFA Deliverable D11.2 “Draft metadata standard for nanoscience data” (M6) for the earlier defined design approach, for relevant information management practices, standards and recommendations, as well as for empirical research done by NFFA JRA3. The deliverable contains a final recommendation for NFFA metadata model, describes effort on its implementation and indicates directions for its further implementation and development.

The deliverable has been discussed and validated through a number of conference calls and electronic communication in JRA3, as well as through engagement with Research Data Alliance groups and external projects outlined in a dedicated section “NFFA metadata interoperability and NFFA engagement with other metadata initiatives”.

Implementation effort is outlined in the “NFFA metadata implementation” section.

Possible choices for potential NFFA metadata model implementation by third parties, as well as a few directions for further work on nano-facilities metadata are discussed in “Metadata operational recommendations and future developments” section.

Particular details of the metadata model design and implementation have been moved into Appendices, in order to keep the main text more concise.

The aforementioned sections “NFFA metadata interoperability and NFFA engagement with other metadata initiatives”, “NFFA metadata implementation” and “Metadata operational recommendations and future developments”, as well as Appendices C, D and E are new to this deliverable. The “Approach and methodology” section is a (much) shortened version of the same from Deliverable D11.2. The “An idealised workflow for NFFA experiments” section and Appendices A, B are inherited and updated from Deliverable D11.2.

1. Approach and methodology

The main purpose of any metadata is satisfying information needs of a certain community. The information needs may be unspecific (common with other communities) or specific for a particular community. From the project delivery perspective, the information needs should be ideally expressed as clearly formulated Use Cases for the existing or proposed information system (IT platform). Therefore to design good metadata both user requirements and IT architecture should be taken into account and in turn should feed considerations for the IT architecture.

For the purposes of metadata design, the project partners' responses to the questionnaire published in the NFFA deliverable D11.2 "Draft metadata standard for nanoscience data" (M6) have been used to identify the existing use cases and practices. For IT architecture considerations, the NFFA deliverable D8.2 "Design of the finalized repository architecture" (M12) has been used, as well as the experience of actual implementation of the NFFA Information and Data Repository Platform (IDRP) [IDRP]. In addition to the bottom-up and in-project considerations, the existing metadata standards and best practices listed in the mentioned NFFA deliverable D11.2 have been taken as a top-down input to the metadata design. Figure 1 illustrates the approach taken in NFFA for the metadata design:

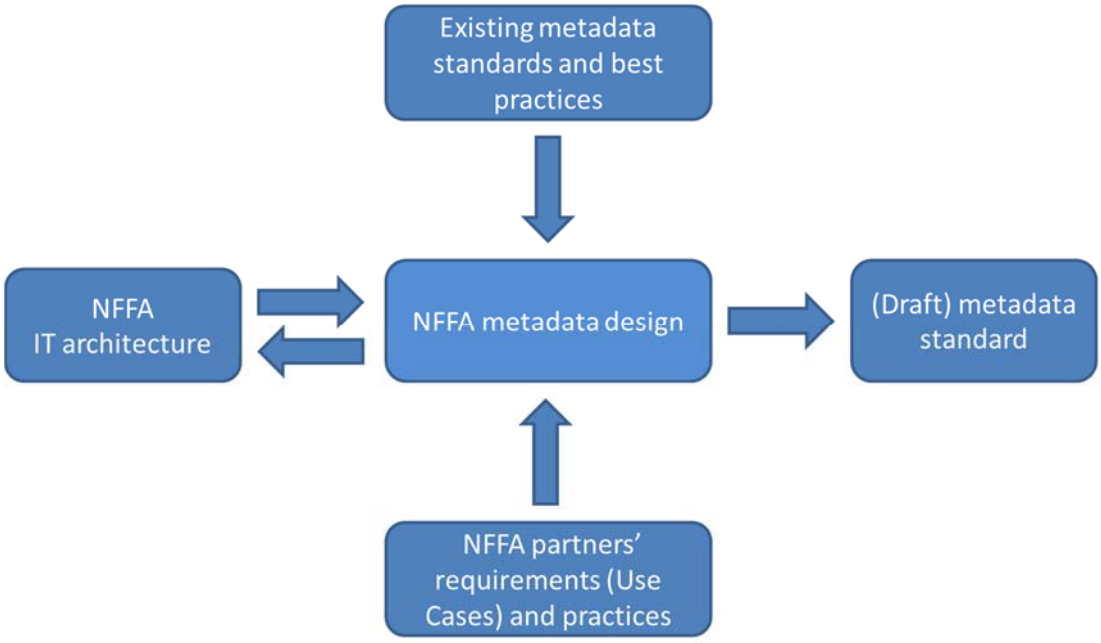

Figure 1: The approach taken in NFFA for the metadata design

The IT architecture, the use cases and practices, and the metadata design can be considered pillars of *enterprise architecture* [EA] that includes both technological and organisational aspects of a (let it be loosely coupled) virtual enterprise that the NFFA project is going to deliver. Therefore, a contribution to the enterprise architecture for nano-facilities experiments (both physical and computational) can be considered a high-level objective for the entire effort of metadata design and implementation.

Existing best practices of information management as well as empirical research through questionnaire disseminated across the NFFA partners have been used as contributions to the development of NFFA metadata. More detailed account of the metadata design approach and methodology is given in D11.2 "Draft metadata standard for nanoscience data" and in conference

papers [NFFA Metadata 1], [NFFA Metadata 2]. An idealized workflow for nano-facilities experiments that the metadata model is supposed to correspond to is outlined in the next section.

2. An idealised workflow for NFFA experiments

Every metadata model describes a certain reality, which in case of the NFFA is the conduct of nano-research experiments (physical or computational) resulted in data assets. We describe an idealized workflow for NFFA experiments and data management that are supposed to be reflected in NFFA metadata. This is necessarily simplified, but gives a sufficient overview of the experimental process to capture the key conceptual entities for the experimental context, which we can describe with the proposed metadata format. The idealized workflow is given in Figure 2.

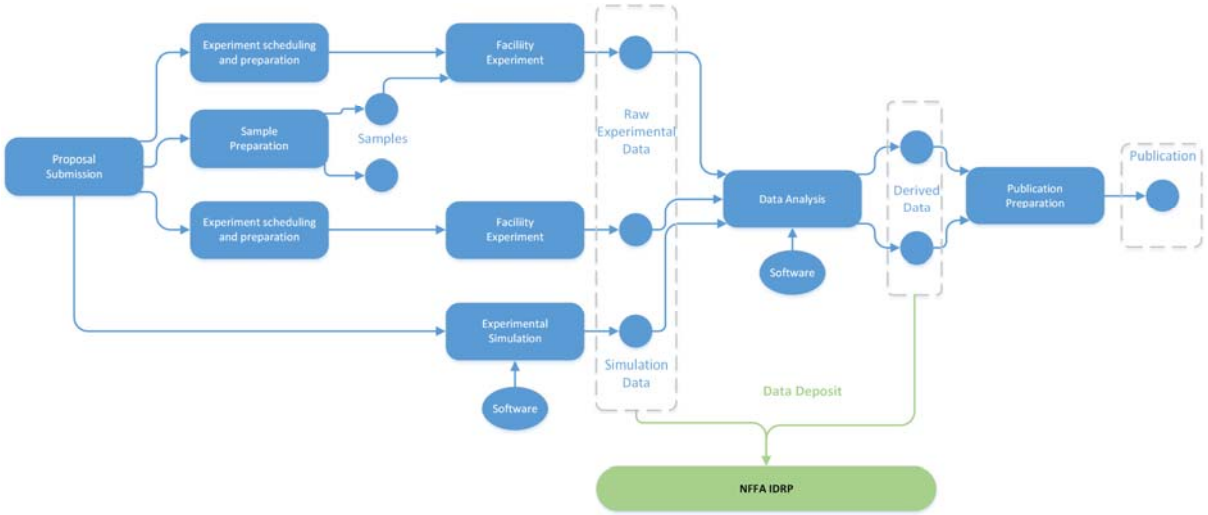

Figure 2: An idealized experimental workflow for NFFA experiments

We briefly discuss the stages in this workflow.

Proposal Submission:	The user submits a proposal to the NFFA submission system, which is peer reviewed. This will specify a number of experiments which are on a number of instruments at facilities, or access to experimental simulation software. If proposal is successful, the proposal can proceed to the next stages.
Experiment scheduling and preparation	Facilities will schedule a visit and allocation of time on instruments. Also, other actions will take place, such as necessary training on the use of instruments, health and safety preparation, handling requirements for samples (which may be for example radioactive, toxic, unstable, or at extreme temperature or pressure).
Sample preparation	Scientists will prepare samples for analysis; note that this may well take a good deal of time. Also, this may require the use of facilities in their

	own right; in this idealized and simplified process description, we omit this, while acknowledging it may take place.
Facility experiment	A sequence of measurements on a set of samples take place on a facility's instrument. These generate raw data sets; we expect the raw data[1] resulting from these measurements to be ingested into the IDRP archive.
Experimental simulation	A sequence of computer simulations of an experiment or model generations are produced by running software packages on computational facilities. These generate simulation data, which again we expect to be ingested into the archive.
Data Analysis	Raw data sets are processed by data analysis software using computational resources. This is likely to be a highly complex process, with multiple runs and iterations, multiple software packages, raw data combined and filtered in different ways, and compared to simulation data; this idealized process gives a highly simplified view. Analysed data sets are the result of this process, which again we expect to be archived.
Publication preparation	The results of the process are interpreted and presented in a documentary form, typically for formal publication, appearing within journals.

At each stage of the process, we should expect to collect metadata describing the context of the experiment at that phase. Thus, this idealized process gives us a framework for identifying appropriate metadata concepts. The concepts themselves have been identified and refined through a series of online and face-to-face meetings of the NFFA JRA3. This resulted in a common vocabulary and entity-relationship diagram presented in Appendix A, and in metadata groups and elements presented in Appendix B.

3. NFFA metadata implementation

The proposed metadata model has been partially implemented by the NFFA Information and Data Repository Platform (IDRP) to provide a generic metadata model for all datasets acquired by NFFA users, either by using instruments or software tools. Due to the way data is acquired in NFFA, there are different phases in which parts of the metadata model can be instantiated:

Proposal Acceptance Phase: All submitted proposals are evaluated following an internal process where they can be rejected or accepted. Before this decision has been made, all proposal information is managed by the NFFA portal. As soon as a proposal gets accepted, the registration of the proposal at the IDRP is triggered automatically. At this point, first parts of the NFFA metadata model can be filled in with basic metadata, namely proposal information and preliminary experiment information.

[1] Or perhaps a "cleanup" or "reduced" version of the raw data, which removes spurious artefacts in the data.

Experiment information have a preliminary nature as during the acceptance phase the facility where a certain experiment may take place is suggested by the principal investigator of the proposal, but not confirmed by the facilities' decision makers, yet.

Experimental Phase: In this second phase, data acquisition takes place at the chosen facilities. Typically, the IDRP is not involved in this phase as data (and 'scientific' metadata) is in most of all cases stored in local data archives, repository systems, or just on a local hard disk.

Post-Experimental Phase: This phase typically starts when the user has left the experiment facility. In this phase, the majority of the remaining metadata entities can be filled in, e.g. experiments, measurements, samples, and data assets. Afterwards, (meta-) data can be accessed via the IDRP. Added value like facetted search, publication, and sharing can be applied to both data and metadata.

In order to register data products acquired in the experimental phase at the IDRP, the principal investigator of the proposal has to trigger the import via the IDRP user interface. In contrast to the initial idea of using a push-based approach for registering data stored in local data archives at the IDRP, the current implementation is realized as pull-based import where the IDRP takes care of mapping external metadata models into the NFFA metadata model. The reason to prefer this approach is the minimization of the effort for other stakeholders, e.g. the users and maintainers of local data archives. Currently, the import of metadata from two data archives has been implemented at the IDRP. The first importer has been implemented in collaboration with EPFL for their Materials Cloud platform [MaterialsCloud]. This platform holds published datasets created using the AiiDA computation platform [AiiDA]. All Materials Cloud entries are publicly accessible and the metadata model is quite compact, thus it could be easily mapped to the NFFA metadata model and appropriately imported to the IDRP.

Table 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

The second implemented use case is the import of CNR-IOM experiments into the IDRP. In this case, an improvement has been done by developing both the pull-based import on the IDRP side, working as for EPFL metadata, and the push-based registration from the local side. The latter has been done by mean of a command line interface able to trigger the ingest of data to the local repository (based on KIT Data Manager repository platform) and to register the corresponding metadata to the IDRP.

Table 2. Crosswalk from CNR-IOM Digital Object to NFFA model.

CNR-IOM Digital Object	NFFA Metadata Model
Investigation.Topic	Experiment.title
Investigation.startDate	Experiment.startTime
Investigation.endDate	Experiment.endTime
Investigation.note	Experiment.experimentDescription
Investigation.uniqueIdentifier	Experiment.experimentIdentifier
DigitalObject.label	Measurement.measurementName
DigitalObject.note	Measurement.measurementDescription
DigitalObject.startDate	Measurement.measurementStart
DigitalObject.endDate	Measurement.measurementEnd
DataOrganizationNode.name	DataAsset.dataAssetName
DataOrganizationNode.size	DataAsset.dataSize
DataOrganizationNode.lastModified	DataAsset.dateOfCollection

In addition, we are in contact with the representatives of the NOMAD repository [NOMAD], but currently it seems to be impossible to map NOMAD metadata to the NFFA model. The main reason is different focus on metadata and the lack of appropriate, public APIs for accessing data and metadata. However, it seems to be rather interesting to employ NOMAD for describing samples and their materials.

In addition to the outlined metadata implementation in IDRP, with samples in a JSON serialized format presented in Appendix C, there is an ongoing work of setting up an interface between IDRP and EUDAT e-infrastructure in order to automate publishing data generated by NFFA experiments in EUDAT B2SHARE service [EUDAT B2SHARE]. The mapping of NFFA metadata elements to EUDAT B2SHARE metadata and a snippet of currently tested JSON serialization of the IDRP record for publishing it in EUDAT B2SHARE are presented in Appendix D. The extensions for these implementations can be informed by a larger collection of metadata elements that are defined in Appendix B.

Another open point that has to be covered in the remaining project lifetime is the support for advanced metadata, e.g. coming from instruments, analysis workflows or the user, as this is naturally not part of the generic NFFA metadata model. However, with the implementation of the generic metadata model in the European data infrastructure for Nanoscience we are able to provide a uniform view on data assets produced in the context of NFFA Europe. This allows us to provide the users with easy-to-use interfaces for obtaining and sharing their scientific results within and outside of NFFA.

4. NFFA metadata interoperability with other metadata initiatives

The NFFA metadata model is aimed at a contextually rich description of nanoscience experiments lifecycle, with data management and data analysis considered to be essential parts of this lifecycle. This defines the position of the NFFA model in respect to other metadata initiatives in nano-research.

CODATA-VAMAS Model

One of such initiatives is a significant effort made by CODATA[2] and VAMAS[3] who established a joint working group for the development of a uniform description system for nanomaterials. The group was international, also multi-discipline with the inclusion of representatives from physics, chemistry, pharmacology, ecology, engineering and other branches of research and technology. Through a number of workshops, the working group developed an elaborated recommendation [CODATA-VAMAS UDS] which, similarly to the NFFA metadata model, does not specify a data format but rather presents a structure of concepts that can be applicable for developing data formats and ontologies, for reporting research results, and for other practical uses.

The main focus of the CODATA-VAMAS model is on a nano-object with the metadata categories (sections) for the description of the object shape, size, physical structure, chemical composition, crystallographic structure and surface description. The model also pays attention to characterization of a collection of nano-objects with the captured concepts of a collection composition, size distribution, association type and topology. The model attempts to address the problem of the nano-objects production and testing, too, describing typical steps involved in those processes.

Cross-walks between NFFA and CODATA-VAMAS models are possible using three NFFA model entities: Sample, Experiment and Measurement. Sample can be related to nano-objects and collections of them in the CODATA-VAMAS model, Experiment can be related to nano-object production steps and Measurement to testsing steps. To enable metadata cross-walks, either the respective entities of the NFFA model can be developed and presented as containers that include metadata definitions copied from CODATA-VAMAS model, or alternatively, these entities can serve as wrappers with pointers to the uniquely identifiable instances defined by the use of CODATA-VAMAS model, so that the respective part of the NFFA model devoted to nano-sample is just an annotation of an external CODATA-VAMAS description.

NOMAD

Another prominent effort has been made by NOMAD (NOvel MAterials Discovery) Laboratory, a European Centre of Excellence (CoE)[4] and is focused on modelling the computation for nanoscience. NOMAD maintains a large repository of input and output files for computer-simulated materials, and

[2] CODATA: Committee on Data for Science and Technology. http://www.codata.org

[3] VAMAS: Versailles Project on Advanced Materials and Standards. http://www.vamas.org/

[4] NOMAD (NOvel MAterials Discovery) Laboratory, a European Centre of Excellence (CoE). https://nomad-coe.eu/

has developed metadata for it.[5] Unlike NFFA where the metadata model has been derived through rounds of communication with nanoscience practitioners and IT architects, the NOMAD approach to metadata design is quite different and can be called opportunistic, as metadata elements are defined looking into the actual results of computational experiments. NOMAD calls this a posteriori approach with the main advantage of it that all significant properties of data can be captured as key-value pairs.

In order to implement this opportunistic or a posteriori approach, NOMAD construct the names of metadata elements on-the-fly depending on the concepts discovered in the results (data output) of a particular computational experiment. In addition to these metadata elements that can be called "topical keys", e.g. "energy total potential" name is a key for the corresponding data value, NOMAD con-sider a hierarchy of descriptors for the runs (executions) of a computer program that are related to particular software configurations, to the results of computation, as well as to theoretical methods used. This gives very context-rich descriptions of the computations actually performed.

Relationship to the NFFA Model

Cross-walks between NFFA model and NOMAD one are possible via the NFFA Experiment entity that can relate to NOMAD "topical keys" that describe a particular experiment, as well as NFFA Measurement entity that can be related to the NOMAD definitions of program runs. The Sample entity of the NFFA model can relate to input data in NOMAD, and Data Asset to the output of NOMAD computation.

The identified cross-walks across the models are compiled in the Table 3.

Table 3. Cross-walks across NFFA, CODATA-VAMAS and NOMAD metadata models

NFFA concept	CODATA-VAMAS concept	NOMAD concept
Experiment	Nano-object production steps	Series of s/w runs
Measurement	Nano-object testing steps	S/w run
Sample	Nano-object or collection of objects	Input data
Data Asset		Output data

It is noticeable that CODATA-VAMAS model is focused on the description of Samples (nano-objects) and the processes directly related to Samples such as production or testing steps but it does not care about data management.

On the opposite, the NOMAD model cares a lot about data; especially about Data Assets resulted from the computational experiment as this data is a source for the extraction of key-value metadata pairs in NOMAD.

Neither CODATA-VAMAS nor NOMAD consider in detail the organizational environment where experiment is conducted, whilst the NFFA model pays a detailed attention to such environment with a few entities like Facility, Proposal or Project catering for this. The data lifecycle in the archive has

[5] NOMAD Meta Info. https://metainfo.nomad-coe.eu/nomadmetainfo_public/info.html

decent means of its description in the NFFA model but not as such in CODATA-VAMAS or NOMAD model.

Overall, the three models have some overlaps which make it possible to specify the above mentioned crosswalks but otherwise the models are complementary to each other. The levels of coverage of a few key aspects of nanotechnology by the three models is presented in Table 4 with the following gradations: Conceptual coverage (where there is at least one concept that can be potentially expanded), Detailed coverage (where there is enough interconnected concepts to cover the aspect) and Unaddressed (when there is nothing or very little in the model to address the aspect).

TABLE 4. Conceptual coverage of nanotechnology experiments by NFFA, CODATA-VAMAS and NOMAD metadata models.

Nanotechnology aspect	NFFA model	CODATA-VAMAS model	NOMAD model
Nano-object	Conceptual	Detailed	Detailed *
Computation	Detailed **	Unaddressed	Detailed
Experiment lifecycle	Detailed	Conceptual	Conceptual
Data lifecycle	Detailed	Unaddressed	Conceptual

*) For in silico (computer simulated) nano-objects only but key-value metadata pairs are potentially applicable to physical objects, too.

**) As all NFFA model concepts are formulated in view of their dual application to physical and computational experiments.

Work within the Research Data Alliance

Apart from the NFFA, CODATA-VAMAS and NOMAD metadata models, other semantic assets such as vocabularies and ontologies can be used to complement or augment the meaning of NFFA metadata concepts or their particular attributes. To better address the metadata interoperability challenge, the collaboratively developed semantic assets should be given a preference before the industry-led specifications.

The Research Data Alliance has established the RDA/CODATA Materials Data, Infrastructure & Interoperability Interest Group[6] and RDA International Materials Resource Registries Working Group (IMRR WG).[7] The NFFA-Europe project has actively participated within these groups.

The IMRR WG has concentrated on developing a distributed registry of resources on materials[8], with an underlying metadata model to capture information on resources within the registry. As a

[6] RDA/CODATA Materials Data, Infrastructure & Interoperability Interest Group. https://www.rd-alliance.org/groups/rdacodata-materials-data-infrastructure-interoperability-ig.html

[7] RDA International Materials Resource Registry Working Group. https://www.rd-alliance.org/group/international-materials-resource-registries-wg/post/rda-materials-registry-working-group

[8] See for example the NIST Materials Resource Registry https://materials.registry.nist.gov/ [DIMA]

metadata model designed for the registration of a wide variety of resources, rather than to capture the experimental context of a data collection, the metadata model is less comprehensive in its design.

Resources collected within these fall within a broad classification of resource types, including Organizations, Data Collections, Datasets, Services, Informational Sites, and Software, and a number of subtypes for each of these. Each resource type then has some additional descriptive fields, mostly from extended Dublin Core. The resource types in the IMRR model thus can map to some of the major entities in the NFFA model: Project and Facility as types of Organization, Data Asset and Data Archive corresponding to Datasets and Data Collection.

The IMRR model includes a notion of "*Applicability*", which provide "Information describing the resource's applicability to a particular domain". As this is instantiated to materials science, the applicability entity includes information on the nature of the material and the experimental methodology used to collect the data. The IMRR WG has then drafted a controlled vocabulary which gives a number of top-level concepts for these applicability concepts and a large number of specific terms across the whole of the materials domain[9]. These controlled terms can then be used to instantiate the attributes from the NFFA model. We give these top-level concepts and how they can be used within the attributes of the NFFA model in Table 5.

Table 5. Top level concepts in the IMRR draft model and usage within the NFFA model

IMRR Vocabulary Concepts	Description	Example	NFFA Attributes
Material types	the category of material studied in the data	nanocomposites, nano-crystalline	Sample Description
Structural features	the primary or prevalent characteristic of the structure of the material of interest	Nanoparticles, nanotubes	Sample Description
Property addressed	a category of property that is sampled by the contained data	Superconductivity, space groups	Measurement description
Experimental methods	the experimental technique used to acquire the data	x-ray absorption spectroscopy, x-ray fluorescence spectrometry	Experimental technique
Computational methods	the computational technique used to acquire the data	reverse Monte Carlo, density functional theory	Experimental technique

[9] https://rd-alliance.org/system/files/documents/Materials_Registry_vocab_draft_170131.pdf

Synthesis and processing	the physical processing or preparation technique applied to the material being studied	chemical vapor deposition, atomic layer deposition	Sample Description

Other work in the RDA is also of interest to the NFFA model. The Metadata IG is developing a generic model describing top-level concepts which can be used as a framework for all metadata models across research. These include notions of Originator, Facility, and Project. The NFFA model should map to these in a straightforward manner.

Another emerging group, endorsed RDA in February 2018, is that for Persistent Identification of Instruments[10] that is particularly relevant for the universal designation of large-scale instruments and facilities involved in the NFFA research lifecycle. A complementary effort in defining consistent references for large-scale instruments and facilities is being made by the ORCID User Facilities and Publications Working Group.[11]

Continuous engagement of the NFFA metadata task with information management community through the RDA allowed to establish a working contact with data practitioners in material science beyond Europe, particularly with National Institute for Materials Science (NIMS) in Japan[12] who are a part of the "Society 5.0" programme promoted by Japanese government as a vision of the future society characterized by the sophisticated integration of cyberspace with physical space.

These contacts with the RDA groups and beyond have served the purposes of validating the NFFA metadata model against wider modelling and practical considerations, and of discovering the cross-walks between the NFFA model and other recommendations to ensure their interoperability, as well as to avoid the duplication of effort where certain parts of metadata, or vocabularies in support of certain metadata elements, can be borrowed elsewhere. This engagement will continue for the rest of the NFFA project so that the NFFA metadata model fitness for purpose and fitness for use will be continuously validated, and the model is linked to other relevant metadata initiatives and recommendations.

[10] RDA Persistent Identification of Instruments group. https://www.rd-alliance.org/groups/persistent-identification-instruments

[11] ORCID User Facilities and Publications Working Group. https://orcid.org/content/user-facilities-and-publications-working-group

[12] National Institute for Materials Science (NIMS). http://www.nims.go.jp/eng/

5. Metadata operational recommendations and future developments

The common vocabulary and the entity-relationship diagram presented in Appendix A define a generic metadata model that reflects the reality of experiments (both physical and computational) in nano-facilities. The "NFFA metadata implementation" section describes the actual effort of implementing parts of the NFFA metadata model in the IT solutions. Apart from this design and implementation effort, operational recommendations are required in order to promote the NFFA metadata model for its wider adoption.

As an example, there are choices of how you aggregate data, e.g. all data files for all samples measured in a particular experiment can be assembled in one package, and then the package is given common descriptions like Facility name, research User name, Data Policy etc. Or, like in the mentioned IDRP implementation and for publishing records of NFFA experiments in EUDAT B2SHARE, one may decide to focus on Measurement as the main artefact to share and publish. However, these choices may not suit actual data management practices or policies of certain Facilities, e.g. they may want to make a Sample rather than an Experiment or Measurement a focal point of their metadata descriptions, or they may want to allow the open publication only of certain data formats (e.g. low resolution images) and publish the more detailed or higher quality data under specific access permissions – then packing up everything related to an Experiment in one package should be avoided.

These operational aspects of NFFA metadata implementation require further discussions and engagement with data practitioners in NFFA participant organizations who will get hands-on experience of working with the IDRP implementation through the remaining time in NFFA project and beyond, and will express their views of how data sharing and data publishing practices, including metadata, should be adopted to their specific needs.

An important operational consideration for NFFA metadata is using it for the records of nano-science experiments (both physical and computational) that are published in common e-infrastructures. The ongoing effort of NFFA metadata publishing in EUDAT e-infrastructure is summarized in Appendix D.

Another operational aspect of NFFA metadata management to be addressed in time remaining in NFFA project and beyond, most likely through the relevant RDA groups mentioned in the "NFFA metadata interoperability and NFFA engagement with other metadata initiatives" section, is raising the quality of NFFA metadata by offering selected controlled vocabularies in support of certain metadata entities like Facilities or Instruments. In some cases, developing new dedicated vocabularies, e.g. publishing the actual NFFA offering in the form of a controlled vocabulary, may be a way to go – and may be a valuable outcome of its own for the NFFA metadata task and its follow-up effort. A possible design of such a vocabulary is indicated in Appendix E, along with considerations for matching vocabulary development effort through RDA.

Correspondences and cross-walks between NFFA metadata model and other metadata models for facilities research, in particular Core Scientific Metadata Model [CSMD] will be worth to explore, as the suggested metadata model is quite generic by design and may have a potential for its direct use

beyond nano-research, or for amendments of existing metadata models. Specifically, the suggested concept of Data Asset that embrace Raw Data (including a result of computer simulation), Analyzed Data, as well as Data Analyses (configurations or/and logs of Data Analyses execution) may have a potential for further elaboration and practical use.

The mentioned operational aspects of NFFA metadata design, implementation and management should be a part of the NFFA project outcomes sustainability considerations.

6. Conclusion

This deliverable draws a baseline under design and implementation of metadata that reflects the business model of nano-facilities characterized by the workflow explained in the “An idealized workflow for NFFA experiments” section and supported by the definitions in the Common Vocabulary (Appendix A).

The main effort of metadata subtask for the reporting period has been focussed on metadata implementation in IDRP, metadata mapping and ongoing experiments on metadata export in EUDAT e-infrastructure, metadata operational recommendations, as well as on problems of the nano-facilities metadata interoperability with other metadata designed and implemented by third parties. The deliverable structure reflects these main directions, with a substantial amount of newly introduced material as well as that partially inherited from the draft deliverable D11.2 as indicated in the “Executive summary” section.

Engagement through RDA, as well as with a wider community of metadata practitioners has been achieved through participation in RDA groups and in conferences. This resulted in peer-reviewed publications [NFFA Metadata 1], [NFFA Metadata 2] which, along with the artefacts in the Appendices A-E, can be considered tangible outcomes of the metadata task.

The remaining project effort of partners involved in the metadata task will be used to address challenges and opportunities outlined in the “Metadata operational recommendations and future developments” section.

References

[AiiDA] Automated Interactive Infrastructure and Database for Computational Science. http://www.aiida.net/

[CODATA-VAMAS UDS] Uniform Description System for Materials on the Nanoscale developed by the CODATA-VAMAS Working Group on Nanomaterials. http://dx.doi.org/10.5281/zenodo.20688

[CSMD] The Core Scientific Metadata Model that supports ICAT data catalogue. https://icatproject.org/user-documentation/csmd/

[DIMA] Dima, A., Bhaskarla, S., Becker, C. et al. Informatics Infrastructure for the Materials Genome Initiative. The Journal of The Minerals, Metals & Materials Society (2016) 68: 2053. https://doi.org/10.1007/s11837-016-2000-4

[EA] Enterprise Architecture: Wikipedia article. https://en.wikipedia.org/wiki/Enterprise_architecture

[EUDAT B2FIND] EUDAT B2FIND data discovery service. https://eudat.eu/services/b2find

[EUDAT B2SHARE] Data sharing service developed by EUDAT project. https://eudat.eu/services/b2share

[IDRP] Information and Data Repository Platform (IDRP). See in D8.2 "Design of the finalized repository architecture" (M12).

[MaterialsCloud] Materials Cloud platform. http://www.materialscloud.org/

[NFFA Metadata 1] V.Bunakov, B.Matthews, T.Griffin, S.Cozzini. Metadata for Experiments in Nanoscience Foundries. In Springer Communications in Computer and Information Science 706 (2017): 248-262. doi: 10.1007/978-3-319-57135-5_18 Open Access version in Zenodo: https://zenodo.org/record/1175958

[NFFA Metadata 2] V.Bunakov, B.Matthews. Metadata for nanotechnology: interoperability aspects. In Metadata and Semantic Research. Communications in Computer and Information Science 755 (2017): 247-252. doi:10.1007/978-3-319-70863-8_24 Open Access version in Zenodo: https://zenodo.org/record/1175964

[NoMaD] NovelMaterials Discoverey Repository, see under http://nomad-repository.eu/cms/

[PANKOS] Proton and Neutron Knowledge Organisation System. http://pan-data.eu/sites/pan-data.eu/files/PaNdataODI-D6.4.pdf

[RDA IMRR WG] RDA International Materials Resource Registry Working Group. https://www.rd-alliance.org/group/international-materials-resource-registries-wg/post/rda-materials-registry-working-group

Appendix A. Common Vocabulary and ER diagram for nano-facilities experiments

Through the iterative discussion within the JRA3, we have identified the major stakeholders, entities and relationships that contribute to the metadata design, and compiled the following Common Vocabulary accompanied by an entity-relationship diagram.[13] This vocabulary is one of the elements of the suggested draft metadata model, and a contribution to what can be called an emerging Enterprise Architecture for nanoscience. The Vocabulary and the ER diagram informed the NFFA implementation of metadata in Information and Data Repository Platform (IDRP), also they can be used by third parties considering their own implementations. The Vocabulary and the ER diagram have been designed and refined in order to conceptually cover both physical and computational experiments.

- **Research User.** A person, a group of them, or an institution (organization) who conduct Experiment on one or more nanoscience Facilities using one or more nanoscience Instruments in order to collect and analyze Raw Data, or is interested in data collected or analyzed by other Research Users on the same or other Facilities. Research Users may have different roles in respect to NFFA Portal.

- **Instrument Scientist.** A person, or a group of them who manage a particular Instrument, or a set of them.

- **Project.** An activity, or a series of activities performed by one or more Research Users on one or more Facilities using one or more Instruments for taking one or more Measurements of one or more Samples during one or more Experiments. Facility, Instrument, Measurement and Sample can refer to computer simulation environment. Project may involve one or more Proposals.

- **Proposal.** An application of Research User for to perform a set of Experiments on one or more Facilities using one or more Instrument.

- **Facility.** An institution (organization), or a division of it that operates one or more nanoscience Instruments for Research Users. For computer simulation, Facility may include hardware or/and software platform or/and services that allow to order and manage computational experiments (so that the software platform serves the purpose of managing software modules that can be considered virtual Instruments).

- **Instrument.** Identifiable equipment (such as a device or a stand or a line) that allows conducting an independent nanoscience research, perhaps without involvement of other Instruments. Instrument is hosted by Facility and used by Research User. Instrument may

[13] All the terms in the Common Vocabulary should be interpreted broadly with the inclusion of “in silico” experimental perspective, even if this is not explicitly mentioned.

be used for Sample production. Measurements conducted on Instrument result in Raw Data in the course of Experiment. Instrument can be in fact a software for computer simulation (a software module or/and a particular configuration of it).

- **Experiment.** Identifiable activity with a clear start time and clear finish time conducted by Research User who uses Instrument to investigate or produce Sample and collects Raw Data about it. Experiment consists of (or includes – in case of Sample production) one or a series of Measurements and may also include one or a series of Data Analyses, potentially specific to Measurements. Experiment can be a computer simulation (computational experiment), or a combination of it with physical Measurements.

- **Measurement.** The act of data collection for a Sample or a series of Samples during Experiment using a particular Instrument. Measurement can be a computer simulation, e.g. a particular run of a program using a particular model, configuration or input. Depending on a particular research context, Measurement may involve measuring the same sample under different conditions, or measuring different samples under the same conditions. Measurement is specific to Instrument: if one has to research the same Sample on a different Instrument it will imply a separate Measurement.

- **Sample.** Identifiable piece of material with distinctive properties (structural, dimensional and others) exposed to Instrument during Experiment. Sample may stand for a model or configuration or data input (or any combination of these) in computer simulation.

- **Raw Data.** Identifiable unit of data collected by Research User during Experiment. Raw Data is a result of Measurement. Unit of data is typically a data file but it can be potentially a data stream, or other form of data relevant in a particular data management context. Raw Data can be a result of computer experiment (simulation). Raw Data is always a part of Data Asset which may bear some semantics of what the data is and the origin/provenance of it.

- **Analyzed Data.** Identifiable unit of data which is a result of Raw Data processing obtained with the use of Data Analysis Software, typically after the end of Experiment. Unit of data is typically a data file but it can be potentially a data stream, or other form of data relevant in a particular data management context. Analyzed Data may or may not be stored in the same Data Archive as Raw Data. If certain software is applied during experiment then a distinction between Raw Data and Analyzed data should be defined by what makes sense in a particular context; as an example, both Raw Data and some Analyzed Data produced on-the-fly may be the outcome of Measurement. Analyzed Data can be a part of Data Asset which may bear some semantics of what the data is and the origin/provenance of it.

- **Data Asset.** A combination of data units which can be Raw Data (including a result of computer simulation), Analyzed Data, or Data Analyses (configurations or/and logs of Data Analyses execution). Depending on a particular data management context, Data Asset can be a dataset, a collection, or other form of data units organization. Data units remain identifiable within Data Asset. Data Asset allows capturing relationships between data units or/and their origin/provenance (e.g. corresponding Measurements or Data Analyses) or/and data curation operations performed on data units (e.g. checksum calculation). Data Asset may also serve as a “container” for different representations (manifestations) of the same

data, e.g. for a collection of semantically equal data files in different formats. Data Asset can be used to express an accumulated result of Measurement (perhaps over multiple Samples).

- **Data Analysis.** The identifiable action of processing Raw Data or/and Analyzed Data, or a Data Asset with Data Analysis Software. Data Analysis can be thought of as something similar to Measurement – just input for it is not Sample but already collected data (raw or/and analyzed or/and contextualized data collections / Data Assets). As Analyzed Data can be a subject of Data Analysis, one can combine Data Analyses in chains or workflows. The definition of workflows and means of modeling them, however, is currently considered to be beyond the project scope, so no specific entities for workflows are introduced in the information model (and resulted metadata model); if someone wants to model workflows, the only means for that is currently Data Asset. Possible relation between Data Analysis and Data Asset is therefore twofold: on one hand, Data Analysis may use Data Assets as input; on the other hand, Data Asset may include Data Analyses configuration (or records of their execution).

- **Data Analysis Software.** Software used for Raw Data analysis (that includes data rendering/visualization) and yields Analyzed Data as an output. If software is used for simulation (computer experiment), is it considered Instrument and should be described as such.

- **Data Archive.** An operational information system (repository) for Raw Data or/and Analyzed Data on a certain Facility with certain rules and principles of data registration and management. Data Archive may or may not be used by Research User(s). Data Archive may include data storage solution (platform, component) and data catalogue solution (platform, component). Term “archive” should be interpreted broadly, i.e. it may be as simple as a file system, also the archive may not be supported by the Facility itself but by a certain third-party that Facility has an agreement with. Data Archive manages Data Assets according to Data Policy (which is perhaps specific to a particular type of Data Asset). Data Archive may be associated with a certain Facility or a group of them, or a certain Instrument or a group of them, or it may be run by a third-party where Facilities or Instruments are willing or obliged to supply their Data Assets (e.g. a discipline-wide or national archive). An example of third-party Data Archive not associated with a particular Facility is EUDAT B2SHARE. NFFA Portal may have one or more Data Archives as a back-end, or interoperate with them.

- **Data Policy.** An identifiable expression of rules and regulations about data management in Data Archive (that includes data ingest) and about data sharing within and beyond Facility. Data Policy may be applicable to Raw Data or/and Analyzed Data. Data Archive may have different Data Policies for different types of Data Assets. NFFA Portal (or its back-end Data Archive) may have one or more Data Policies, too.

- **Data Manager.** Identifiable person, a group of them, an organizational unit, or a machine agent (software) who operate Data Archive on a certain Facility or in the third-party establishment that Facility or NFFA Portal have an agreement with. Having a clear identity and clear description of Data Manager is important for managing data harvesting (or federated data infrastructure) in NFFA Portal and resolving potential issues with Data Policies. It is also important for planning, performing and monitoring Data Curation Activities. Data

Managers may have different roles; more than one role may be required by Data Archive or NFFA Portal, e.g. with different sets of permissions.

- **Data Curation Activity.** An identifiable unit of work performed by Data Manager (in a certain role), or by a few of them. Activity can be data ingest, data integrity check, data transformation, restructuring or annotating data or collections of them, or anything else. Data Curation Activity is performed on Data Assets according to Data Policies.

- **NFFA Portal.** An IT service for nanoscience data discovery and sharing; the service may include one or more than one of: Graphical User Interface; Application Programming Interface; data ingestion and data publishing feeds; data sharing, data annotation and data analysis components. NFFA portal is used by Research Users and is underpinned by Data Archives in participating Facilities. Research Users may be registered with NFFA Portal. Data Archives of participant organizations may interact and interoperate with NFFA Portal – both technically and organizationally, e.g. by having Service Level Agreements for data supply in NFFA Portal.

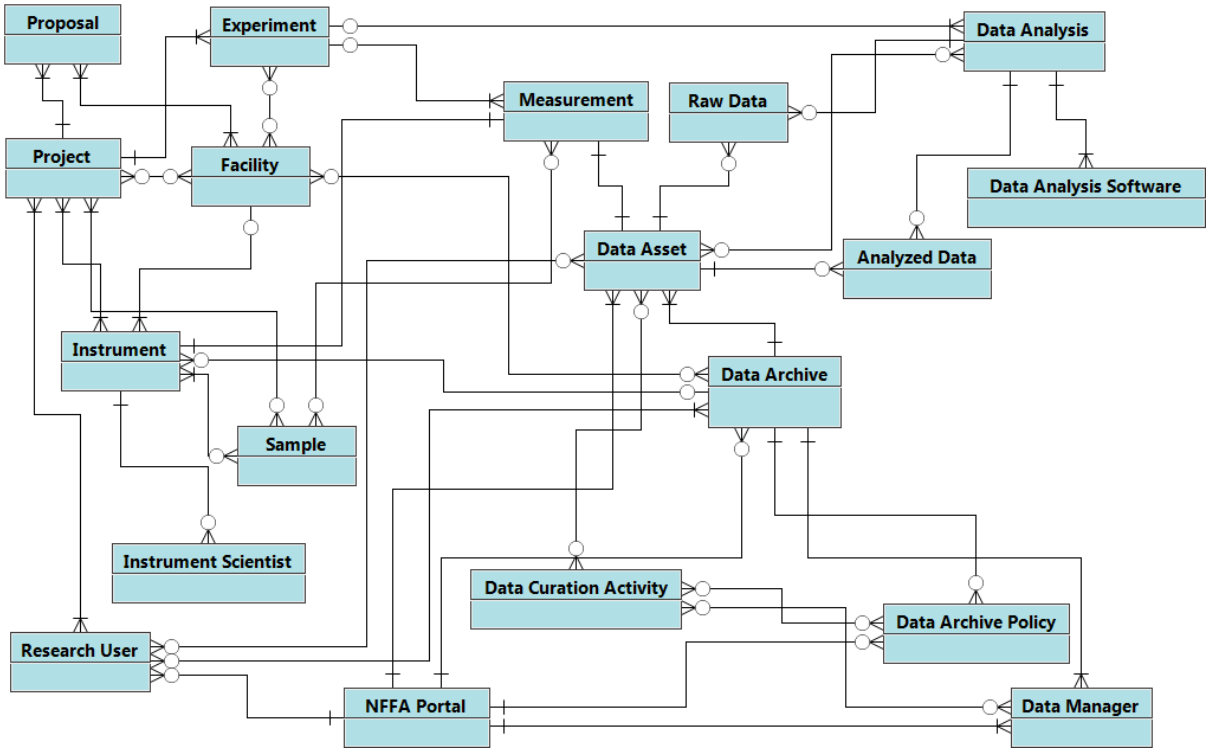

Figure 3: Entity-Relation Diagram for the Common Vocabulary

To illustrate the entities in the Common Vocabulary and their relationships we give an E-R diagram on Figure 3. The diagram can be used for IT Architecture design; however its main purpose is having a clear graphical representation of major information entities defined in the Vocabulary, and relations between them.

Appendix B. Metadata groups and elements

The metadata "building blocks" (groups of metadata elements) are presented in Figure 4. They roughly follow the stages of data acquisition: planning and conducting experiment, capturing and structuring data, then ingesting them in the NFFA data archive. Also, they are likely to be supplied by different groups of actors with different specialisms: experiment description – by researchers themselves, perhaps with the participation of the nano-facility User Office (administrative unit), data assets description - by automated tools and IT staff, NFFA record wrapper – by Data Managers.

These metadata groups and elements informed the NFFA implementation of metadata in Information and Data Repository Platform (IDRP), also they can be used by third parties considering their own implementations.

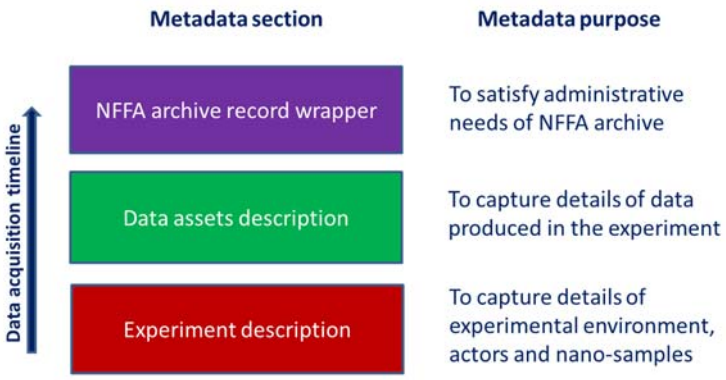

Figure 4. Metadata groups of elements and their purpose.

The suggested metadata elements are grouped by the information entity in the common vocabulary they describe. These information entities are related to the information needs identified above. The Table 6 below summarizes the groups of metadata elements suggested; colours of the cells correspond to the metadata groups in Figure 4.

Table 6. Information needs of NFFA portal users (including the portal administrative staff).

Information entity	Ingest data	Manage data (within NFFA portal)	Disseminate data	Find data	Identify data	Obtain data
Research User			Y	Y	Y	Y
Instrument Scientist	Y	Y				

Project			Y	Y	Y	Y
Proposal	Y	Y				
Facility	Y	Y	Y	Y	Y	Y
Instrument			Y	Y	Y	
Experiment			Y	Y	Y	
Sample			Y	Y	Y	
Data Asset	Y	Y	Y	Y	Y	Y
Raw Data	Y	Y	Y	Y	Y	Y
Analysed Data	Y	Y	Y	Y	Y	Y
Data Analysis	Y	Y			Y	
Data Analysis Software	Y	Y			Y	
Data Archive	Y	Y				Y
Data Manager	Y	Y				Y
Data Policy	Y	Y				
NFFA Portal		Y		Y		

Metadata elements are then defined within these groups. Table 7 details essential elements in the metadata groups, with their suggested value type and description. In subsequent work on the metadata metadata design we will relate these elements to suitable existing standards and recommendations discussed above for compatibility and integration with existing data sources.

Some elements are designated as mandatory elements that all NFFA partners should supply when submitting (or registering) a data asset in NFFA portal. The mandatory fields are suggested so that we have enough coverage for all information needs, also in some cases the decision about making a field mandatory is ingrained in the definition of a respective information entity; as an example, we suggest making the time when Experiment ended a mandatory element as Experiment is by definition an activity having clear boundaries in time (so having at least one boundary recorded will help with the Experiment identification).

Further elements are designated as unique when there is expected to be at most one value for a given information entity. If this is not so defined the element can occur multiple times for a particular information entity. If an element is declared optional it may be omitted.

For some information entities, having both an ID (which can be internal – specific to the facility or data management platform) and a PID (which should be universal) is suggested: one of them intended for managing data in the NFFA software platform, and another for publishing the project outcomes beyond its boundary and lifespan.

Table 7. Essential elements in the metadata groups, with their suggested value type and description

Metadata elements and subgroups	Related Information entity	Value Type	Cardinality	Description
User ID	Research User	Identifier	Mandatory; Unique	Unique identifier for the user
User name	Research User	Text	Mandatory	Commonly user name of the user
User Identifier	Research User	Text	Optional	Unique identifier (PID/URI) assigned to the user by an external organization e.g. ORCID.
User affiliation	Research User	Text[14]	Optional	Institutional affiliation of the user.
Instrument Scientist ID	Instrument Scientist	Identifier	Mandatory; Unique	Unique identifier for the Instrument Scientist
Instrument Scientist name	Instrument Scientist	Text	Mandatory	Commonly user name of the Instrument Scientist
Instrument Scientist Identifier	Instrument Scientist	Text	Optional	Unique identifier (PID/URI) assigned to the Instrument Scientist by an external organization e.g. ORCID.
Instrument Scientist affiliation	Instrument Scientist	Text	Optional	Institutional affiliation of the Instrument Scientist.
Project ID	Project	Identifier	Mandatory, Unique	Unique identifier for the Project. It may be assigned by a funding body (so known prior to applying for the facility time slot), or the project ID generation feature/service can be offered by a proposal registration system
Project name	Project	Text	Mandatory	Name for the project
Project description	Project	Text	Optional	Textual description of the project

[14] We may wish to extend this to an institution entity at a later stage.

Proposal ID	Proposal	Identifier	Mandatory, Unique	Unique identifier for the Proposal
Proposal description	Proposal	Text	Optional	Textual description of the project
Facility ID	Facility	Identifier	Mandatory, Unique	Unique identifier for the Facility
Facility Identifier	Facility	Text	Optional	Unique identifier (e.g. PID/URI) assigned to the Facility by an external organization.
Facility name	Facility	Text	Mandatory	Common name for the Facility
Instrument ID	Instrument	Identifier	Mandatory, Unique	Unique identifier for the Instrument
Instrument Identifier	Instrument	Text	Optional	Unique identifier (e.g. PID/URI) assigned to the Instrument by an external organization.
Instrument name	Instrument	Text	Mandatory	Common name for the Instrument
Instrument type	Instrument	Text	Mandatory	Class of instrument, e.g. diffractometer, spectrometer. Should be term from a known controlled list.
Experimental technique	Instrument	Text	Optional	Class of experimental technique supported by instrument, e.g. diffraction, spectroscopy. Should be term from a known controlled list. Note that instruments can support more than one technique,
Experiment ID	Experiment	Identifier	Mandatory, Unique	Unique identifier for the Experiment
Experiment Identifier	Experiment	Text	Optional	Unique identifier (e.g. PID/URI) assigned to the Experiment by an external organization, for example a visit identifier assigned by a Facility.
Experiment title	Experiment	Text	Mandatory	Short descriptive title for the Experiment which is how the experiement is commonly referred.

Experiment start time	Experiment	DateTime	Optional	Date and time (if relevant) the period of the experiment started
Experiment end time	Experiment	DateTime	Mandatory	Date and time (if relevant) the period of the experiment was completed.
Experiment description	Experiment	Text	Optional	Descriptive abstract of the experiment
Measurement ID	Measuremen t	Identifier	Mandatory; Unique	Unique identifier for the Measuremnt
Measurement Type	Measuremen t	Tag/Text	Optional	Designation of the Measurement purpose or/and character, as an example: * Calibration * Data collection on sample * Simulation * Physical measurement There may be more than one type/tag associated with the measurement, e.g. the Measurement can be simulation AND data collection on (virtual) sample in the same time. Whether tags are coming from controlled vocabulary or are arbitrary can be left for the facility to decide. Measurement Type can be used for the Measurement record consistency checks, e.g. there may be a local (facility-specific) requirement that Measurements which are not Calibrations require Sample ID presence in the Measurement record (whilst in general a reference to Sample remains optional).
Measurement Name	Measuremen t	Text	Optional	Name of the measurement
Measurement Description	Measuremen t	Text	Optional	Description/explanation of the measurement
Measurement Start	Measuremen t	Time	Optional	Time when the measurement started

Measurement End	Measurement	Time	Optional	Time when the measurement ended
Sample ID	Sample	Identifier	Mandatory, Unique	Unique identifier for the Sample. This is unique for the *specific instance* of the sample
Sample Identifier	Sample	Text	Optional	Identifier (e.g. PID/URI) assigned to the Sample within an external identification scheme. For example, this could be a IUPAC International Chemical Identifier (*InChI*), or a reference in a standard chemical database.
Sample name	Sample	Text	Mandatory	Commonly used name for the sample. Could refer to terms in controlled vocabularies for materials.
Sample description	Sample	Text	Optional	Textual description of the sample
External metadata reference	Sample	URL	Optional	Reference for more detailed metadata for a detailed description of sample characteristics in a domain specific metadata format.
Data ID	Raw Data	Identifier	Mandatory, Unique	Unique identifier for the Raw Data object
Data Identifier	Raw Data	Text	Optional	Identifier (e.g. PID/URI) assigned to the Raw Data within an external identification scheme or data management system
Data name	Raw Data	Text	Mandatory	Filename or stream name for the Raw Data Object
Data format	Raw Data	Text	Optional, Unique	Format of the Raw Data. Should refer to a standard name for the format in a controlled vocabulary.
Data format Identifier	Raw Data	Text	Optional	Identifier for the data format as assigned by an external organization.
Data type	Raw Data	Text	Optional	Type of the data in the Raw Data object

Data size	Raw Data	Integer	Optional, Unique	Size of the Raw Data object in Bytes.
Data checksum	Raw Data	Integer	Optional, Unique	Calculated checksum of the Raw Data object
Date of collection	Raw Data	DateTime	Mandatory, Unique	Date and time if relevant of the completion of the collection of the Raw Data object
Intellectual property rights	Raw Data	Text	Optional	Licencing information or other IPR rights
Data ID	Analysed Data	Identifier	Mandatory, Unique	Unique identifier for the Analysed Data object
Data Identifier	Analysed Data	Text	Optional	Identifier (e.g. PID/URI) assigned to the Analysed Data within an external identification scheme or data management system
Data name	Analysed Data	Text	Mandatory	Filename or stream name for the Analysed Data Object
Data format	Analysed Data	Text	Optional, Unique	Format of the Analysed Data Object. Should refer to a standard name for the format in a controlled vocabulary.
Data format Identifier	Analysed Data	Text	Optional	Identifier for the data format as assigned by an external organization.
Data size	Analysed Data	Integer	Optional, Unique	Size of the Analysed Data object in Bytes.
Data checksum	Analysed Data	Integer	Optional, Unique	Calculated checksum of the Analysed Data object
Date of creation	Analysed Data	Date Time	Mandatory, Unique	Date and time if relevant of the completion of the collection of the Analysed Data object
Intellectual property rights	Analysed Data	Text	Optional	Licensing information or other IPR rights
Software ID	Data Analysis Software	Identifier	Mandatory Unique	Unique identifier for the Software object

Software package name	Data Analysis Software	Text	Mandatory	Commonly used name for the software package
Software version	Data Analysis Software	Text	Optional	Specific version number of the software package.
Software package identifier	Data Analysis Software	URL	Optional	Link to more information on the software package, additional metadata, downloadable packages
Data Archive ID	Data Archive	Identifier	Mandatory Unique	Unique identifier for the Data Archive
Data Archive Name	Data Archive	Text	Mandatory	Name of data archive that supplied data in NFFA (that may not be equal to the name of facility)
Data Archive reference	Data Archive	Text	Optional	Dereferenceable identifier (e.g. PID/URL) assigned to the Data Archive
Data Archive description	Data Archive	Text	Optional	Description of the archive ownership or/and scope or/and mode of operation
Data Archive keywords	Data Archive	Text	Optional	Keywords for the Data Archive - for registration in the external information systems, and for search engines.
Data Manager ID	Data Manager	Identifier	Mandatory; Unique	Unique identifier for the Data Manager
Data Manager name	Data Manager	Text	Mandatory	Commonly user name of the Data Manager. This may be a machine agent name if “archivist” is a software
Data Manager reference	Data Manager	Text	Optional	Dereferenceable identifier (PID/URI) assigned to the Data Manager by an external organization e.g. ORCID.
Data Manager affiliation	Data Manager	Text	Optional	Institutional affiliation of the Data Manager
Data Policy ID	Data Policy	Identifier	Mandatory Unique	Unique identifier for the Data Policy

Data policy type	Data Policy	Text	Optional	Type of data policy that can be: general data management policy, data integrity checks policy, data release policy, or any other policy relevant to a particular part of the data lifecycle. One of the policy types to consider can be SLA seen as an expression of “data ingest policy” or “data supply policy”, or other sort of policy that defines a contract between different legal or organizational entities
Data policy description	Data Policy	Text	Optional	Policy expressed in a textual form. This can be used if no dereferenceable identifier for the policy is available
Data policy reference	Data Policy	Text	Optional	Dereferenceable identifier (e.g. PID/URL) assigned to the
Portal ID	NFFA Portal	Identifier	Mandatory Unique	Unique identifier for the Portal
Portal version	NFFA Portal	Text	Optional	Version of the portal
Portal keywords	NFFA Portal	Text	Optional	Keywords for the Portal - for registration in the external information systems, and for search engines.
Portal contact information	NFFA Portal	Text	Optional	Contact details for the institution and person assigned as contact for the portal.

Appendix C. NFFA metadata serialization in IDRP

Appendix A with a common vocabulary and entity-relationship diagram and Appendix B with metadata groups and elements can be considered metadata design artefacts that are universal for use in NFFA and by third parties with a similar research workflow.

Particular serialization of the suggested metadata model in certain formats is left to IT implementers to decide upon. This Appendix contains snippets of JSON serialization for particular metadata objects and elements as implemented in the NFFA Information and Data Repository Platform (IDRP) [IDRP]. These can serve as examples for the third parties who are willing to develop their own implementations of the same generic metadata model, in JSON or other serialization format.

Serialization of Proposal metadata in IDRP


```
{
  "proposalId": "string",
  "proposalTitle": "string",
  "projectId": "string",
  "members": {
    "userId": "string",
    "userName": "string",
    "userEmail": "string",
    "userAffiliation": "string"
  },
  "principalInvestigator": {
    "userId": "string",
    "userName": "string",
    "userEmail": "string",
    "userAffiliation": "string"
  },
  "proposalDescription": "string",
  "embargoUntil": "2018-02-14T09:13:25.282Z",
  "registrationTime": "2018-02-14T09:13:25.282Z"
}
```

Serialization of Experiment metadata in IDRP


```
{
  "proposalId": "string",
  "experimentId": "string",
  "experimentIdentifier": "string",
  "experimentTitle": "string",
  "startTime": "2018-02-14T09:13:25.468Z",
  "endTime": "2018-02-14T09:13:25.468Z",
  "experimentDescription": "string"
}
```


Serialization of Instrument metadata in IDRP


```
{
  "facilityId": "string",
  "instrumentId": "string",
  "instrumentIdentifier": "string",
  "instrumentName": "string",
  "instrumentType": "string",
  "experimentalTechnique": "string"
}
```

Appendix D. NFFA metadata publishing in EUDAT

NFFA metadata publishing in EUDAT e-infrastructure[15] is a work in progress with two main considerations:

- Automated publishing of metadata from Information and Data Repository Platform (IDRP) [IDRP] in EUDAT B2SHARE service [EUDAT B2SHARE] which in turn will be automatically re-publishing these records in EUDAT B2FIND service (data catalogue) [EUDAT B2FIND].
- Providing the NFFA research community (users) with an ability to assign NFFA-specific metadata when the users decide to publish their data directly in EUDAT B2SHARE using its graphical user interface (with further automated metadata promotion from EUDAT B2SHARE to EUDAT B2FIND).

Working on the implementation of the first consideration is naturally opportunistic and depends on whatever metadata records are actually available through the NFFA proposal system and through particular data collection systems in participating facilities. From the EUDAT B2SHARE point of view, there are two parts of metadata record: common fields imposed by EUDAT B2SHARE service (with community-specific values for them) and community-specific metadata fields. Table 8 presents common fields with suggested NFFA-specific values for them and Table 9 presents community-specific fields and possible sources for them.

Where Table 9 indicates a Facility as a source of metadata field value, this means that this value can be supplied from IDRP in the NFFA implementation, in cases where IDRP captures data records from a participating Facility. Other (non-NFFA) implementations can use suggestions in Table 9 as a guidance for publishing nano-facilities data records in common e-infrastructures.

Table 8. EUDAT B2SHARE common metadata fields and their NFFA-specific values.

MD field	Type	Explanation	Decision made (recommended value for the field)
Community	string (required)	Identifier of the community to which the record has been submitted	"NFFA.eu"
Title	string (required)	The title of the uploaded resource - a name that indicates the content to be expected.	Measurement title
Description	string	A more elaborate description of the resource. Focus on a description of	Measurement description

[15] EUDAT e-infrastructure. https://www.eudat.eu/

		content making it easy for others to find it and to interpret its relevance quickly.	
Authors	array <string>	The record author(s).	"IDRP" in case the metadata record is published automatically; whoever submits the record in EUDAT B2SHARE in case this is done via EUDAT B2SHARE GUI
Open Access	boolean (required)	Indicate whether the resource is open or access is restricted. In case of restricted access the uploaded files will not be public, however the metadata will be.	Default: TRUE (as NFFA intends to publish only un-embargoed data in EUDAT B2SHARE)
Licence	string	Specify the license under which this data set is available to the users (e.g. GPL, Apache v2 or Commercial). Please use the License Selector for help and additional information.	Skipped in case of automated publishing from IDRP (as there is currently no recommended default licence for NFFA Open Access data); whatever is chosen by the submitter in case of individual sharing via the EUDAT B2SHARE GUI
Keywords	array <string>	A list of keywords that characterize the content.	Used instrument(s) and a proposal ID in case of automated publishingfrom IDRP; whatever is chosen by the submitter in case of individual sharing via the EUDAT B2SHARE GUI
Contact Email	[email]	Contact email information for this record	IDRP instance manager in case of automated record submission; whatever is chosen by the submitter in case of individual sharing via the EUDAT B2SHARE GUI
Discipline	string	The scientific discipline linked with the resource.	"Nanoscience"

Embargo Date	[date-time]	The date marking the end of the embargo period. The data stored in the record will become publicly available on the specified date at midnight. The record metadata is always public.	Not needed as we only put un-embargoed data into B2Share
Contributors	array <string>	The list of all other contributors. Mention all persons that were relevant in the creation of the resource.	Research User names from the proposal, or at least Principal Investigator
Resource Type	array <string>	The type of the resource.	“RAW” or “ANALYSED” depending on what kind of data is being submitted
Alternate identifier	string	Any kind of other reference such as a URN, URI or an ISBN number.	No actual recommendation
Version	string	Denote the version of the resource.	No actual recommendation
Publisher	string	The entity responsible for making the resource available, either a person, an organization, or a service.	“IDRP” in case of automated submission; whatever is chosen by the submitter in case of individual sharing via the EUDAT B2SHARE GUI
Language	string	The name of the language the document is written in	No actual recommendation

Table 9. EUDAT B2SHARE community-specific metadata fields and sources for their population with values.

MD field	Type	Explanation	Sources of the field value	Comment
Proposal description	String	Name of the research proposal	NFFA portal application form fields (concatenated): "Title", "Abstract", "State of the art" and "Objectives". Alternative source: proposal description in IDRP.	
Project name	String	Name of the project that benefits from the experiment	Project name if mentioned in the NFFA portal application form	There is no simple way to reliably extract the project name from the research proposal (application form). Also this is currently not a requirement to mention the name of the project in an NFFA proposal.
Facility name	String (required)	Name of the site where the experiment is conducted	NFFA portal application form "Preferred sites" field. This is only a user request though; the actual site offered to the user may be different.	The actual site offered to the user may be different to what the user requested in the application form – or be one of a few for which the user has applied. So names of the actual sites offered are best controlled by facilities themselves, and be supplied by facilities.

Instrument name	String	Instrument used for the experiment	Facility where the experiment has been actually conducted	From an IT point of view, the actual source much depends on a software platform used for data collection at the facility
Experiment ID	String	Identifier for the experiment	Facility where the experiment has been actually conducted	Supplied by a particular facility
Experimental techniques	Array <string>	Experimental techniques used for the experiment	NFFA portal application form "New step" field.	1) Experimental techniques actually used may be different from those requested in the application form. The actually used are best collected from facilities where the experiments have been conducted. 2) Experimental technique is considered an Instrument attribute in the current metadata model – and the only one that NFFA proposal cares about. The rest of the Instrument information may come only from facilities where experiments have been actually conducted.
Sample ID	String	Sample identifier	Facility where the experiment has been actually conducted	User can provide descriptions of more than one sample in the application form but how (and whether at all) facilities are going to

				identify samples is unclear and may vary across facilities. Yet we will need some form of ID in order to group Sample substance/formula/physical state
Sample description	String	Description of a sample	NFFA portal application form fields (concatenated): “Sample substance”, “Sample chemical formula”, “Sample physical state”, “Sample size”, “Sample hazards”	
Measurement type	Designation for the experiment purpose or character	String	Facility where the experiment has been actually conducted	Sourced from the facility
Measurement ID	Identifier for the measurement	String	Facility where the experiment has been actually conducted	Sourced from the facility; can be a facility experiment “run” ID
Measurement name	Name for the measurement	String	Facility where the experiment has been actually conducted	Sourced from the facility; can be used in the absence of standardized measurement IDs

Measurement description	Description / explanation of the measurement	String	Facility where the experiment has been actually conducted	Sourced from the facility
Measurement start	Time when the measurement started	String (with a format agreed)	Facility where the experiment has been actually conducted	Sourced from the facility
Measurement end	Time when the measurement ended	String (with a format agreed)	Facility where the experiment has been actually conducted	Sourced from the facility

The JSON snippet below presents a partial implementation for metadata from Table 6 that is under testing now; it is centered around Measurement attributes collected at a particular NFFA facility (which may be a computational platform in case of an in silico experiment). In the actual (Production) implementation this should be complemented with common metadata from Table 8 (at least with mandatory fields, most of which can be static / literals), as well as with metadata from Table 9 that are sourced from the NFFA proposal system and.

Serialization of Measurement metadata in the interface between IDRP and EUDAT B2SHARE

{

"experimentId": "idrp-test-experiment-1",

"measurementType": "RAW",

"measurementId": "d3224a07-10d1-48af-ace2-79a48b969bb4",

"measurementName": "First measurement",

"measurementDescription": "This is the first measurement stored at the IDRP. The next step to test will be adding data to this measurement.",

"measurementStart": "2017-02-09T00:00:00Z",

"measurementEnd": "2017-02-09T00:00:00Z",

"embargoUntil": "2017-02-10T09:20:12Z",

"sampleId": null,

"instrumentId": null,

"globalPid": "https://trng-b2share.eudat.eu/records/20813e9fe5444840ace9725d0e270f71"

}

Appendix E. Controlled vocabularies for raising NFFA metadata quality

Quality of metadata can be improved if metadata elements are filled in with values from controlled vocabularies. Two vocabularies have been evaluated for their use in the NFFA metadata: material types vocabulary being developed by the RDA International Materials Resource Registry Working Group [RDA IMRR WG] and Proton and Neutron Knowledge Organisation System (PANKOS) vocabulary [PANKOS] .

The RDA material types vocabulary is a work in progress led by NIST[16] where NFFA-associated researchers contributed and plan to contribute more. This may allow to consistently describe materials referred from Sample metadata element. This vocabulary though should be used only opportunistically as it is going to be quite generic (categorizing all sorts of materials, not nano-materials specifically), also a material is just one attribute of Sample, with much more attributes (ideally, referring to controlled vocabularies) required to sensibly describe nano-samples.

The PANKOS vocabulary (ontology) was an output of PaNdata collaboration[17] and is again not specifically targeted at nano-research but can be used to describe the NFFA nano-facilities offering for research users. The RDF-XML snippet below presents an example of NFFA facility offering in PANKOS format (for DESY reflectometer):

```
<Ontology xmlns="http://www.w3.org/2002/07/owl#"
    xml:base="http://www.purl.org/pankos"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xml="http://www.w3.org/XML/1998/namespace"
    ontologyIRI="http://www.purl.org/pankos">
   <Prefix name="" IRI="http://www.purl.org/pankos#"/>
   <Prefix name="owl" IRI="http://www.w3.org/2002/07/owl#"/>
   <Prefix name="rdf" IRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>
   <Prefix name="xml" IRI="http://www.w3.org/XML/1998/namespace"/>
   <Prefix name="xsd" IRI="http://www.w3.org/2001/XMLSchema#"/>
    <Prefix name="rdfs" IRI="http://www.w3.org/2000/01/rdf-schema#"/>
```

[16] National Institute of Standards and Technology by the US Department of Commerce. https://www.nist.gov/

[17] PANdata - the Photon and Neutron data infrastructure initiative. http://pan-data.eu/

```
    <Prefix name="pankos" IRI="http://www.purl.org/pankos#"/>
   <Declaration>
      <NamedIndividual IRI="#DESY"/>
    </Declaration>
   <Declaration>
      <NamedIndividual IRI="#DESY_Reflectometer"/>
    </Declaration>
    <ClassAssertion>
      <Class IRI="#Facility"/>
      <NamedIndividual IRI="#DESY"/>
    </ClassAssertion>
    <ObjectPropertyAssertion>
      <ObjectProperty IRI="#inFacility"/>
      <NamedIndividual IRI="#DESY_Reflectometer"/>
      <NamedIndividual IRI="#DESY "/>
    </ObjectPropertyAssertion>
    <ObjectPropertyAssertion>
      <ObjectProperty IRI="#hasInsrument"/>
      <NamedIndividual IRI="#DESY"/>
      <NamedIndividual IRI="#DESY_Reflectometer"/>
    </ObjectPropertyAssertion>
    <ObjectPropertyAssertion>
      <ObjectProperty IRI="#supportsTechnique"/>
      <NamedIndividual IRI="#DESY_Reflectometer"/>
     <NamedIndividual IRI="#X-RayDiffraction"/>
    </ObjectPropertyAssertion>
</Ontology>
```