Metadata for Experiments in Nanoscience Foundries

Vasily Bunakov, Tom Griffin, Brian Matthews

& Stefano Cozzini

Published version information

Citation: V Bunakov et al. "Metadata for Experiments in Nanoscience Foundries." In: Kalinichenko L., Kuznetsov S., Manolopoulos Y. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2016. Communications in Computer and Information Science, vol. 706. Springer (2017): 248-262.

doi: 10.1007/978-3-319-57135-5_18

The final publication is available at Springer via http://doi.org/10.1007/978-3-319-57135-5_18

Metadata for Experiments in Nanoscience Foundries

Vasily Bunakov[1], Tom Griffin[1], Brian Matthews[1], Stefano Cozzini[2]

[1] Science and Technology Facilities Council, Harwell, Oxfordshire, UK
[2] Istituto Officina dei Materiali, CNR, Trieste, Italy
{vasily.bunakov, tom.griffin, brian.matthews}@stfc.ac.uk
cozzini@iom.cnr.it

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

1 Introduction

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

Metadata design is a part of a joint research activity within NFFA-EUROPE that takes empirical input from the project participants, and also takes into account state-of-the art standards and practices. Metadata design is an incremental effort of the project; this work presents the first stage resulting in a high-level metadata model that is agnostic to the actual data management situation in participating organizations yet is able to capture significant features of physical and computational nanoscience experiments.

Compared to the well-known metadata recommendation for nanoscience developed by CODATA-VAMAS Working Group On the Description of Nanomaterials [7] which is heavily focussed on nano-samples description, the metadata model we are developing in NFFA-EUROPE is intended to well reflect the lifecycle of data collected in nanoscience experiments (both physical and computational), and then archived for the purposes of further data discovery and data sharing. This is why this model makes the most sense for data practitioners in nanoscience and for research users who want to discover and explore the context of data assets resulted from nanoscience experiments.

This work adds to the earlier published effort of metadata design for nanoscience [13]. It expands on the motivation for the development of a new metadata model for nanoscience, details metadata implementation effort, specifically the ongoing work of metadata crosswalks between NFFA-EUROPE and EUDAT [8] (in Section 3.4), and presents a new refined version of the Common Vocabulary (in Appendix A) that underpins all metadata design and relates to other metadata artefacts that constitute the high-level metadata model. Also, this work outlines the identified challenges of metadata design and suggests directions for its further development (in Section 4).

2 Approach and Methodology

2.1 General Approach

The major purpose of any metadata is satisfying information needs of a certain community. "Community" should be understood in broad terms and includes machine agents, to ensure human-to-human, human-to-machine and machine-to-machine interoperability.

The information needs may be generic (common with other communities) or specific for a particular community. From the implementation point of view, the information needs should be expressed as clearly formulated Use Cases for the existing or proposed information and data management systems (IT platforms), so that the role of metadata in the data workflow can be clearly identified. A good metadata design should take into account user requirements and IT architecture, and in turn should feed considerations into the IT architecture. Figure 1 illustrates the approach taken in NFFA-EUROPE for the metadata design.

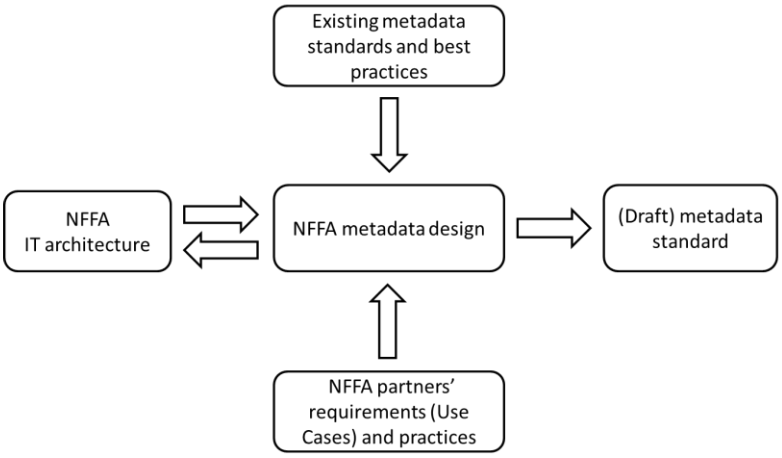
Existing metadata standards and best practices
NFFA IT architecture
NFFA metadata design
(Draft) metadata standard
NFFA partners' requirements (Use Cases) and practices

Fig. 1. Approach to NFFA-EUROPE metadata design.

Metadata can be considered a part of the *enterprise architecture* that includes both technological and organizational aspects of a loosely coupled virtual enterprise that the NFFA-EUROPE project is going to deliver for the European nanoscience community.

The metadata design then represents one of the pillars of the enterprise architecture design of the NFFA-EUROPE virtual enterprise, the other two pillars being business analysis and IT architecture design. Working on all three pillars should be mutually communicated and eventually aligned, which allows for the delivery of a quality enterprise architecture.

A good practice of information and data management adopted in the NFFA-EUROPE context is getting a good common understanding shared by the project partners about what actors (stakeholders), entities and relationships are most important in their domain and hence should be taken into account for the metadata design, and what are less important or too specific to be taken into common consideration. Through the iterative discussions in the project, we picked up the most relevant Roles and Responsibilities in the nanoscience domain, and mixed them up with the major Entities definitions that often constitute a basis of a structured formal knowledge representation (ontology) of a certain subject domain, but in our less ambitious case will form a basis of a reasonable metadata schema.

These discussions resulted in the Common Vocabulary (see in Appendix A) which is a concise Body of Knowledge that describes information entities and relations between them that are most common in the project partners' experimental and data management environment. As a particular although again generic representation of this Body of Knowledge, we have described this in an Entity-Relationship (ER) diagram (see in Section 3.2).

The Common Vocabulary and the ER diagram taken together with metadata groups and elements (see in Section 3.1) constitute a generic metadata model and a baseline for all discussions about NFFA-EUROPE metadata. They are the basis for the detailed metadata model with the definition of metadata elements and relations among them. The detailed metadata model, when agreed upon, can be further represented in a certain serialisation format such as XML, RDF, or JSON. There is an early indication driven by technology considerations that a detailed master representation of NFFA-EUROPE metadata will be in JSON format.

The practice of iterative metadata development which we follow in NFFA-EUROPE has already got then a sound foundation – a Common Vocabulary, ER diagram and practical suggestions on metadata groups and elements – with the detailed metadata design and its particular (serialised) representations to be elaborated in later stages of the project.

2.2 Top-Down Input: Relevant Information Management Frameworks

The case for metadata collection and use can be specific to nanoscience, yet there are general information needs that are typical for a wide variety of users and that have been developed in other branches of science and information management.

One of the mature information design frameworks is Functional Requirements for

Bibliographic Records (FRBR) [2] that considers four basic information needs (user tasks) in regards to information: “Find”, “Identify”, “Select” and “Obtain”. The ultimate goal is of course getting the information resource, yet between searching for it and obtaining it, the resource should be identified as the one being sought, and selected as being useful for the user [1]. Each task may involve certain subtasks, e.g. selection may require checks on the resource context and on its relevance to the actual user’s needs.

Another elaborated information design framework of relevance is the Reference Model for an Open Archival Information System (OAIS) [3], a widely-known functional model for long-term digital preservation. If expressed in terms of information practitioner needs (user tasks) similarly to FRBR, the OAIS basically deals with three categories of them: “Ingest (into archive)”, “Manage (within archive)” and “Disseminate (from archive)”. Each of these tasks may be complex and involve a number of interrelated subtasks, e.g. managing information in the archive may imply provenance and integrity checks, managing access to information, and administration / reporting.

Overall, the OAIS framework should be able to provide a good coverage of what NFFA-EUROPE needs to consider for sensible data collection, archiving and provision towards the end users (researchers in nanoscience), and the FRBR framework should be able to cover the end user needs for information retrieval. The respective areas of coverage and user categories relevant to NFFA-EUROPE are illustrated by the following table:

Framework (a source of best practices)	**OAIS**	**FRBR**
General use case	Data collection, management and dissemination	Data retrieval
User categories	Data archives administrators IT specialists	End users (nanoscience researchers)
Information needs (user tasks)	Ingest data Manage data Disseminate data	Find data Identify data Select data Obtain data

Table 1. Information management frameworks and their coverage of NFFA-EUROPE scope.

Being general in nature, OAIS and FRBR are still able to provide good recommendations for NFFA-EUROPE practices of information and data management. In particular, OAIS emphasizes the need of having a clear agreement between the data producer and the archive, and a clearly defined format for data exchange between them – so called Submission Information Package, whilst FRBR emphasizes the importance of having a clear identity for data assets.

2.3 Bottom-Up Input: Questionnaire Responses and Common Vocabulary

A questionnaire was used to collect the NFFA-EUROPE partners’ responses about their data management practices and most popular data management solutions. The questionnaire inquired on the following aspects of data management in nano-facilities:

- Intensity of experiments and of resulting data flow
- Popular data formats
- Data catalogue software
- Data catalogue openness
- Data management policy
- Metadata standards for data catalogue
- Persistent identifiers for data
- User management platform
- Popular third-party databases and information systems

In total, seventeen responses out of the twenty project partners were received and reviewed. They showed very different levels of data management maturity. From the responses, the following priorities for metadata design were identified:

- One experiment to many samples and one sample to many data files relationships should be supported.
- A common set of metadata fields for data discoverability should be agreed upon, possibly based on an existing popular standards or recommendation for data discovery.
- User roles with different permissions for access to metadata should be developed. This means the metadata model will need to represent users as well as data.
- It is reasonable to develop a common data management policy for NFFA-EUROPE, or a set of policies with different flavours of access to data.
- Having links to external reference databases is valuable to ensure the high quality of metadata yet this will mean additional effort so should be de-scoped from the initial design of metadata.

In addition to the questionnaire where responses were collected from research offices or relevant research programme representatives, a common vocabulary of terms and definitions relevant to nanoscience data management was compiled and then refined by the IT teams of participating NFFA-EUROPE organizations (see in Appendix A). The vocabulary contains commonly agreed terms with definitions; it serves as a basis for the design of information entities (groups of metadata elements) and contributes to the earlier mentioned NFFA-EUROPE “virtual enterprise” architecture.

2.4 Side Input: IT Architecture Considerations

As an additional consideration for principal metadata design, we used the draft of NFFA-EUROPE Data System Architecture that defines the outline design of the NFFA-EUROPE portal, which considered the generic use case of the same user performing a measurement on multiple facilities. Generic use cases when one user wants to access data produced by another user, or wants to release data into the public domain are currently not being considered. These may be considered in future, so should be taken into account within an extensible metadata design.

The draft architecture suggests that data should be harvested from individual facilities in a suitable "packaged" format, with METS [6] as a potential candidate as it supports the provision of descriptive, administrative, structural and file metadata. For the descriptive part of metadata, the purpose of having the data assets discoverable is emphasized in the draft architecture. For the administrative metadata, the importance of intellectual property information and information about the data source (provenance) is emphasized. For the structural metadata, having the information about the organization, perhaps structured in a hierarchical way, is suggested. For the file metadata, having the list of files that constitute a digital object (data asset) and having pointers to external metadata files are deemed most important.

After considering the draft of Data System Architecture, the conclusion was that we could take METS as "the role model" metadata standard that informs us about good practices of metadata design but we should not accept it as a default universal solution, as it does not cover all information needs of NFFA-EUROPE users. As to particular elements of metadata suggested by the Data System Architecture draft, the fields for capturing intellectual property information and provenance are easily most important ones as they affect the data assets reusability that should be one of the important outcomes of the NFFA-EUROPE project.

3 Implementation

3.1 Metadata Groups and Elements

The top-down, bottom up and side requirements resulted in the basic structure of the proposed metadata model that is illustrated by Figure 2. This metadata structure generally reflects the data lifecycle in nano-science: first, an experiment is planned and conducted; it then results in some data assets (which can be measurements performed during nano-sample characterization, or controlled parameters of the sample physical production or computer simulation), then the archive that holds data assets should have its own operational requirements – again reflected in the respective section of metadata.

The suggested metadata elements are presented as a matrix in Table 2 to make explicit the coverage of identified information entities (Common Vocabulary terms) and of earlier identified information needs (categories of them, see Section 2.2). Certain elements are in common with the Core Scientific Metadata Model [4] already in use in some of the facilities involved in the NFFA-EUROPE project.

Mandatory and optional metadata fields (attributes) for each element were defined and shared amongst project participants for further discussion in the form of the project deliverable [5]. Some elements and attributes of them were further refined through the process of mapping NFFA-EUROPE metadata to the metadata scheme used in EUDAT B2SHARE service [9], [10] which is detailed in section 3.4.

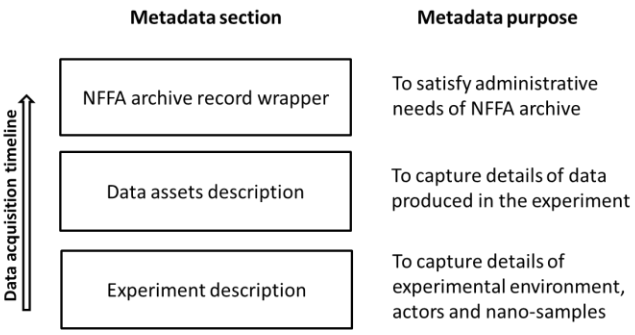
Metadata section
Metadata purpose
Data acquisition timeline
NFFA archive record wrapper
To satisfy administrative needs of NFFA archive
Data assets description
To capture details of data produced in the experiment
Experiment description
To capture details of experimental environment, actors and nano-samples

Fig. 2. Metadata groups of elements and their purpose.

Metadata section	Information entity	Ingest data	Manage data	Dissemi-nate data	Find data	Identify data	Obtain data
Experiment description	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 assets description	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	
Archive record wrapper	Data Archive	Y	Y				Y
	Data Manager	Y	Y				Y
	Data Policy	Y	Y				
	NFFA-EUROPE Portal		Y		Y		

Table 2. Metadata elements and information needs coverage.

3.2 Entity-Relationship Diagram

As a basis for further, more detailed metadata design and as a contribution to the IT architecture design, the Entity-Relationship diagram presented by Figure 3 has been agreed.

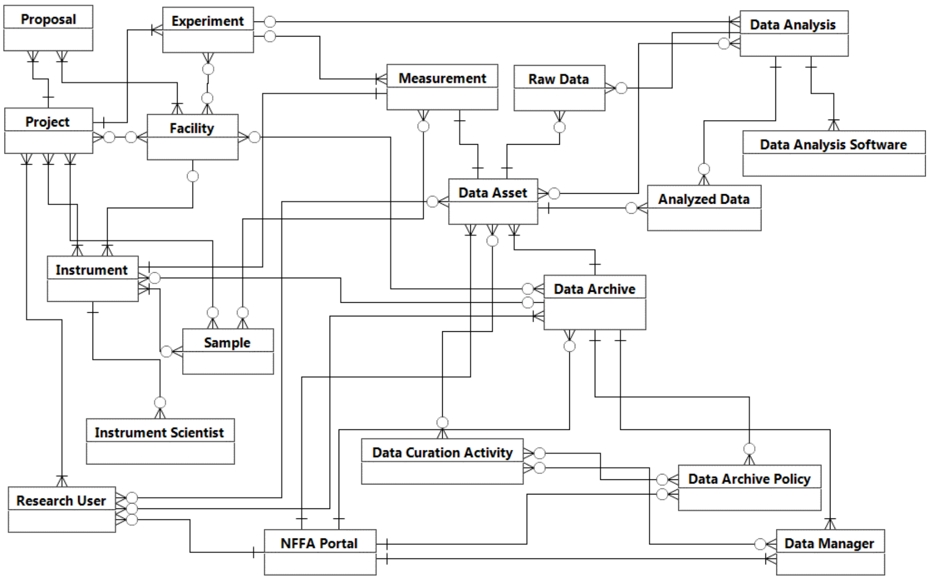

Fig. 3. NFFA-EUROPE metadata entity-relationship diagram

This ER diagram has proven to be a useful tool for all discussions about NFFA-EUROPE metadata design; the entities in it relate to the terms in the Common Vocabulary (see in Appendix A). The diagram allows at least three different perspectives: Research User-centric, Facility-centric, or Data Archive-centric, which reflects the natural data lifecycle in NFFA-EUROPE when the user first submits a research proposal, then conducts the actual (physical or/and computational) experiment, then NFFA-EUROPE takes the resulted data in custody.

3.3 Metadata Operational Recommendations

The metadata elements suggested are not all we need for having a successful metadata framework in NFFA-EUROPE. In addition, there should be established metadata management practices, ideally assisted by clear recommendations for NFFA-EUROPE partner organizations of how to assign and curate metadata.

For example, there are choices of how you aggregate data: let us say 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 such as Facility name, research User name, Data Policy etc. However, this 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 a focal point of their metadata descriptions.

Another operational aspect important for the NFFA-EUROPE metadata scheme adoption by nanoscience community is actual levels of metadata that users and nano-facilities will be happy to provide when submitting research proposals and conducting (physical or computational) experiments. Initial evaluation performed using the research proposals submission system that is already in operation [11] has shown that it can provide satisfactory amount of metadata for Research User, Sample, Project and Proposal Entities. More metadata values for Facility, Instrument, Instrument Scientist, Experiment and Measurement entities should be supplied either by facilities or by users in the time of the actual experiment. The rest of metadata elements will be filled in with actual values by NFFA-EUROPE data portal. The population of metadata scheme with the actual values will be happening thus by various stakeholders and in stages that can be designated as "Research proposal submission" – "Experiment" – "Data archiving".

These operational aspects of NFFA-EUROPE metadata implementation will require further engagement and discussions with data practitioners in NFFA participant organizations.

3.4 Publishing NFFA-EUROPE Data Records in EUDAT Research Infrastructure

EUDAT project [8] supported by the European Horizon 2020 programme delivers common services in support of research data management and research data processing. EUDAT collaborates with other European projects that favour using the EUDAT services or software in place of development of their own functionally similar services or software.

NFFA-EUROPE have decided on the pilot use of EUDAT B2SHARE software platform [9], [10] in order to publish the data resulting from NFFA-EUROPE experiments in nano-facilities. The publication of NFFA-EUROPE data in B2SHARE will be subject to a data policy that is currently under the development in NFFA-EUROPE; in the meanwhile, there is a collaborative effort in NFFA-EUROPE and EUDAT to develop a metadata crosswalk from NFFA-EUROPE to EUDAT B2SHARE schema.

There is a common part of metadata schema in EUDAT B2SHARE that is universal across all communities who are using B2SHARE, and there is a community-specific part that B2SHARE platform can adopt as a template and then offer it for all individual researchers or institutions in the respective research domain – which will be nano-science in our case.

Both the universal part of NFFA-EUROPE metadata for B2SHARE and the community-specific part are first being discussed within NFFA-EUROPE, then with EUDAT representatives, to ensure semantic interoperability of metadata elements. The universal (community-unspecific) part of NFFA-EUROPE metadata crosswalk to B2SHARE schema is now fully agreed, and the crosswalk for the community-specific part of metadata is under development.

The actual data publishing from NFFA-EUROPE data portal to EUDAT B2SHARE instance will be performed using the B2SHARE API. We also foresee the situation when individual nano-science researchers or institutions will like to augment the B2SHARE instance (prepopulated with the automatically acquired data) with their own data uploaded via the B2SHARE user interface. Either a bundle of data and its metadata, or metadata only (with a reference to the corresponding data asset) can be uploaded in B2SHARE, that will give a proper flexibility for the nano-science researchers to share their data according to their local policies and personal preferences.

4 Identified Challenges and Further Developments

Apart from the clearly perceived need to develop, in addition to metadata schema, some operational recommendations for metadata curation (see in Section 3.3), much needs to be done about better identity of metadata elements and values of their attributes.

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) has been suggested: one of them intended for managing data in the NFFA-EUROPE software platform, and another for publishing the project outcomes beyond its boundary and lifespan.

It is the project's intention to get a registered URI for each metadata element – using PURL.ORG or similar services for managing namespaces and unique identifiers. The exact service and naming will be agreed through a dedicated discussion in the project. Unique URIs for metadata elements can constitute a basis for the further sharing of nanoscience data records as Linked Open Data, although the actual implementation of it is going to be beyond the NFFA-EUROPE scope.

Another addition likely to be required will be specifically designed fields for cross-linking metadata elements. As an example, Instrument may require a field, or a few, as a "foreign key" (which is only a metaphor, as the actual metadata representation may not be relational-based) to Facility; the same applies to a desirable link between Proposal and Research User, as well as to a number of other cases. The exact design of these fields dedicated to cross-linking of metadata elements will depend on the chosen format/syntax for metadata serialization: XML, RDF, JSON, or anything else.

We consider the necessity of introducing roles or types for certain metadata elements, up to the point of convergence of certain metadata elements into more universal ones supplied with a role or type attribute (a tag). Prime candidates for this would be Raw Data and Analyzed Data elements, as both during and after the experiment, it may make sense to deal with „data continuum“ where the data is assigned with approproate tags depending on particular data collection, filtering or analysis steps.

Also the detailed design of Data Asset has been postponed, as it will be heavily driven by the IT Architecture considerations and the pilot implementation of data portal, initially with only a few participating nano-facilities. A preliminary discussion suggested that METS could be a good metadata recommendation to model Data Asset, or to serve as a conceptual wrapper to the bespoke Data Assest modeling.

Certain considerations have been given to the notion of data processing workflows, although owing to the conceptual and technological complexity of workflows they are left beyond the metadata design in NFFA-EUROPE. Some suggestions of how one could model workflows, to a certain extent, by the means of the suggested NFFA-EUROPE data model can be found in Common Vocabulary (see Appendix A, specifically the definition of Data Analysis).

For Sample, there is a reserved metadata attribute for linking a brief record of it to a detailed one that is formed according to an existing standard. CODATA UDS [7] is considered a good candidate for a detailed and well-structured description of nanoscience samples, so the current vision is just to rely upon a rich description of nano-samples offered by CODATA UDS if the NFFA-EUROPE ever identifies a need for a detailed samples description. The promotion of this or other suitable metadata standard for samples will be done then through the engagement effort across the project partners; this effort should be more of an operational nature rather than immediately related to the task of NFFA-EUROPE metadata design.

The Working Groups and Interest Groups of Research Data Alliance [12] are considering appropriate metadata frameworks for data sharing, both domain-focussed, e.g. dedicated to materials science, and cross-domain like those considering the best practices for persistent data identifiers. This is complementary to the approach of NFFA-EUROPE, and we foresee that this will be an appropriate forum for the continued metadata design for nanoscience.

5 Conclusion

The process of metadata development in NFFA-EUROPE so far has produced an agreed common approach with its mapping to the existing metadata frameworks and best practices. It has defined the common vocabulary, the structure of metadata groups and elements, the provisional list of mandatory and optional attributes, and the ER diagram that can be used both in metadata design and in IT architecture design. The high-level metadata model will be further refined through project work in NFFA-EUROPE and through discussions in the wider nanoscience community, with cooperating e-infrastructures like EUDAT and with relevant Research Data Alliance groups.

Acknowledgements

This work is supported by Horizon 2020 NFFA-Europe project www.nffa.eu under grant agreement no. 654360.

Appendix A. Common Vocabulary for Nanoscience Data Management

This vocabulary is one of the components of the suggested high-level metadata model, along with the metadata groups and elements (see in Section 3.1) and ER diagram (see in Section 3.2) and hence as explained in section 2.1 it is a contribution to the NFFA enterprise architecture, with a specific role of giving a common terminology for data practitioners in nanoscience. All the terms should be interpreted broadly with the inclusion of “in silico” experimental perspective, even if this is not explicitly mentioned. The vocabulary will be modified and expanded as necessary through further project works on metadata.

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 User may be assigned with a role, e.g. to designate the user as a principal investigator.

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 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. 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 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 beyond the project scope, so no specific entities for workflows have been introduced in the 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. Examples of Data Curation Activity: data ingest, data integrity check, data transformation, restructuring or annotating data or collections of them. 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.

References

[1] Hider, P.: Information resource description: Creating and managing metadata. Facet Publishing, London (2012)

[2] Functional Requirements for Bibliographic Records (FRBR). Final Report. http://archive.ifla.org/archive/VII/s13/frbr/

[3] Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book). Issue 2, June 2012. CCSDS (The Consultative Committee for Space Data Systems), Washington DC (2012)

[4] The Core Scientific Metadata Model (CSMD). https://icatproject.org/user-documentation/csmd/

[5] Draft metadata standard for nanoscience data. NFFA project deliverable D11.2. February 2016.

[6] METS: Metadata Encoding and Transmission Standard. http://www.loc.gov/standards/mets/

[7] Uniform Description System for Materials on the Nanoscale, Version 2.0. CODATA-VAMAS Working Group On the Description of Nanomaterials. 25 May 2016. doi:0.5281/zenodo.56720

[8] EUDAT e-infrastructure project. http://www.eudat.eu/

[9] EUDAT B2SHARE service. https://b2share.eudat.eu/

[10] EUDAT B2SHARE user documentation. https://eudat.eu/services/userdoc/b2share

[11] NFFA User Guide. http://www.nffa.eu/apply

[12] Research Data Alliance. http://www.rda.org/

[13] Bunakov, V., Matthews, B., Griffin, T., Cozzini, S.: Metadata for Nanoscience Experiments. In: Selected Papers of the XVIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016). CEUR Workshop Proceedings, Vol-1752. urn:nbn:de:0074-1752-7