Metadata for nanoscience experiments

© Vasily Bunakov © Tom Griffin ©Brian Matthews
Science and Technology Facilities Council,
Harwell, Oxfordshire, UK
vasily.bunakov@stfc.ac.uk tom.griffin@stfc.ac.uk brian.matthews@stfc.ac.uk
© Stefano Cozzini
Instituto Officina dei Materiali
Trieste, Italy
cozzini@iom.cnr.it

Abstract

Metadata is a key aspect of data management. This paper describes the work of NFFA project on the design of a metadata standard for nanoscience community. 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 support a single entry point for research proposals supported by the project, 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.

The project requires setting up the IT infrastructure for managing research proposals and substantial amounts of data resulted from physical and computational experiments. A common metadata model that supports different stages of the nanoscience research lifecycle is essential to unified researchers' experience across locations, and also for the design and operation of IT infrastructure components.

Metadata design is a part of a joint research activity within NFFA that takes empirical input from the project participants, 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 nanoscience physical and computational experiments.

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

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 project perspective, the information needs should be expressed as clearly formulated Use Cases for the existing or proposed information and data management systems (IT platforms). A good metadata design should take into account user requirements and IT architecture, and in turn should feed considerations for the IT architecture.

The IT architecture, the use cases and practices, and the metadata design can be considered pillars of *enterprise architecture* that includes both technological and organizational aspects of a loosely coupled virtual enterprise that the NFFA project is going to deliver for the European nanoscience community.

The main purpose of metadata design effort in NFFA project can be formulated then as giving the adequate support for that widely defined enterprise architecture for nanoscience. This has an implication of metadata design from "first principles", i.e. by pondering over existing best practices of information management, use cases for nanoscience and information technology opportunities (and limitations) rather than adopting any existent metadata standard.

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 [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 mature information design framework of relevance is the Reference Model for an Open Archival Information System [3], a popular 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 the archive)", "Manage (within the 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 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 are illustrated by the following table:

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

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

Being general in nature, OAIS and FRBR are still able to provide good recommendations for NFFA 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 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 20 project partners were received and reviewed. They showed very different levels of data management maturity. From the responses, the following priorities the 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, 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 organizations ([5]). The vocabulary contains about twenty 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 "virtual enterprise" architecture.

A particularly important use case to be supported by the metadata model should be the situation when the same researcher (or a research group) applies for experimental time on more than one facility – as the nature of experiment may require this – yet the researcher wants a seamless experience across nanoscience facilities, with a single entry point for data management.

Another conclusion based on responses to the questionnaire is that computational experiments in nanoscience become common and can be mixed up with physical experiments, so there should not be an artificial division between the two.

2.4 Side input: IT architecture considerations

As an additional consideration for principal metadata design, we used the draft NFFA Data System Architecture that defines the outline design of the NFFA portal, which considered the generic use case of the same user performing a measurement on two different 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 architecture, the conclusion was that we could take METS as "the role model" metadata standard for data packaging that corresponds to a specific entity in the NFFA generic metadata model – Data Asset. As to particular elements of metadata suggested by the IT 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 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 1.

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. Mandatory and optional metadata fields (attributes) for each element were defined and shared amongst project participants for further discussion ([5]).

3.2 Entity-relationship diagram

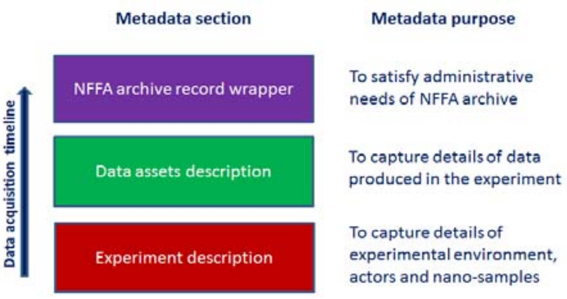
Metadata section
Metadata purpose
Data acquisition timeline
NFFA archive record wrapper
To satisfy administrative needs of NFFA archive
Data assets description
To capture details of data produced in the experiment
Experiment description
To capture details of experimental environment, actors and nano-samples

Figure 1 Metadata groups of elements and their purpose.

Table 2 Metadata elements and information needs coverage.

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		

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 2 has been agreed.

3.3 Metadata operational recommendations

The metadata elements suggested are not all we need for having a successful metadata framework in NFFA. In

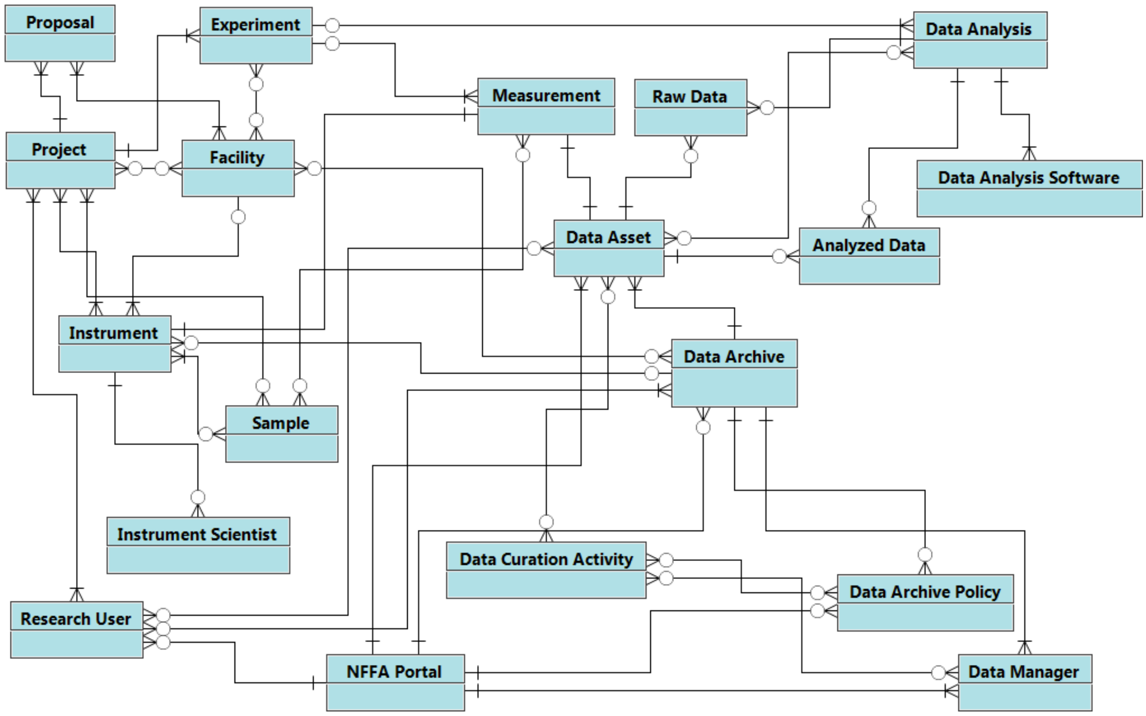

Figure 2 Entity-Relationship diagram for NFFA high-level metadata model.

addition, there should be established metadata management practices, ideally assisted by clear recommendations for NFFA 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 like 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.

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

4 Conclusion

The NFFA metadata development so far has produced an agreed common approach with its mapping to the existing metadata frameworks and best practices. It has defined a common vocabulary, 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 and through discussions in the wider nanoscience community. Also the state-of-the-art metadata development for nanoscience that may cover specific entities in our generic metadata model, e.g. CODATA UDS [7] for Sample, should be looked into in more detail, to see the opportunities for mutual mapping and cross-walks between different metadata models.

Acknowledgements

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

References

[1] Philip Hider. Information resource description: Creating and managing metadata. Facet Publishing, 2012.

[2] Functional Requirements for Bibliographic Records (FRBR). Final Report. http://archive.ifla.org/archive/VII/s13/frbr/ Retrieved 20 May 2016.

[3] Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book). Issue 2, June 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf Retrieved 20 May 2016.

[4] The Core Scientific Metadata Model (CSMD). https://icatproject.org/user-documentation/csmd/ Retrieved 20 May 2016.

[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] COADATA UDS: Uniform Description System for Materials on the Nanoscale http://www.codata.org/uploads/Uniform_Description_System_Nanomaterials-Published-v01-15-02-01.pdf