The ABC metadata model

Evaluation of the Amsterdam Biomedical Concise metadata model

RESEARCH DATA MANAGEMENT, RESEARCH SUPPORT, AMSTERDAM UMC

From: Joost Daams and Evelien van der Schaaf-de Wolf
Version: 1.0
Date: 15-05-2023

Introduction

In the world of biomedical research, managing research data is of great importance to keep an overview over all different resources. Resources can refer to any object used in research practice, such as a data set, images, software or an entire research project.

Storing, preserving, and sharing resources must be done carefully to ensure the reliability and validity of results and to promote FAIR research (Findable, Accessible, Interoperable, Reusable). To support this metadata models are available. However, existing metadata models are too generic.

For this reason, there is a need for a metadata model that can link generic standards and specific metadata generated by researchers and systems. The goal of this model is not only to connect both, but also to describe all FAIR elements of biomedical research in a concise way. Being concise is important, because of minimalizing work load for researchers and the heterogeneity of applications and RDM systems. Therefore the metadata model is not a ready to use system, but a checklist of items with clear concept definitions.

This corresponds to the following objectives of research data management (RDM): "The objective of research data management is to ensure reliable verification of results, facilitate data security and minimize the risk of data loss, enable research continuity through secondary data use, and increase research efficiency. Research data management is an integral part of the research process and helps make research as efficient as possible. The key objectives (...) are to provide capability to store, preserve and access increasingly large volumes of electronic research data. Additionally, by following a good data management strategy, researchers will meet the requirements set out in their institution's RDM policy." [perplexity https://www.perplexity.ai/?s=c&uuid=57fa23c3-cfeb-47a0-a0c4-e1ca350f4c3a, last consulted 8-2-2023]

The deliverables of this project are:

- A metadata model with a concept definition for each item;
- Mapping of the items of the model to Dublin Core, DataCite, and DCAT standards;
- A machine-readable JSON format to facilitate universal implementation;
- An overview of what is still missing from the model (such as synthetic data, and descriptions of software development projects).

Provenance, audit trails and linking systems (e.g., via API) are outside the scope of this project, but they are connected to the implementation of the metadata model.

Methods

Within the expert panel "Data matters" of Amsterdam UMC, a first version of the model was developed. We wanted to verify this first version by evaluating the metadata model for missing items, clarity and relevance.

There is a distinction between usability and the model itself. Usability mainly depends on the implementation and clarity of the items. In this project, we wanted to evaluate the model's content. Usability comments and implementation tips that we received were described separately in the results section.

We consulted with Amsterdam UMC methodology department for advice on how to test the model, and a protocol was developed. It was suggested that we survey both junior and senior researchers. In addition we surveyed several data managers/specialists.
We created a **survey** using the tool LimeSurvey, in which we reviewed the metadata model with the respondents and collected their feedback. Questions asked for each item in the metadata model were:

- Is this item clear?
- Is this item relevant?
- Do you miss any items?

Furthermore there was the possibility to provide additional comments. See "Appendix 1_info_metadataschema_20211112.pdf" for the invitation send to the participants and "Appendix

2_Metadataschema_DataDict_LS_PROD_20211012.pdf" for the data dictionary of the LimeSurvey project.

Additionally, we held three separate **expert sessions** to assess whether the model would remain intact in a particular implementation of granularity:

1. Creating metadata on the (e)CRF level, for example a data dictionary;
2. Mapping metadata on various levels of a medical image catalogue (currently under development) using the XNAT data model (https://wiki.xnat.org/documentation/how-to-use-xnat/understanding-the-xnat-data-model);
3. Creating a FAIR description of bioinformatics laboratory research (https://bioinformaticslaboratory.eu/).

In the meantime we participated in an **implementation project** testing the metadata model in practice. One of the objectives of this project was the development of a machine readable adaptation of the metadata model in JSON format. For this purpose the project employed the file management system iRODS/YoDa to

- Create a web interface for describing items to be published on the publication platform DataverseNL;
- Exchange (meta)data between platforms (i.e. DataverseNL and myDRE).

A new version of the metadata model was developed based on input from both the survey and the expert sessions. We created a definition and/or description of each item based on the feedback of the respondents. We reused existing definitions as much as possible. An item mapping was made to ensure alignment with the Dublin Core, DataCite and DCAT standards.

Results

Survey

In total 19 respondents reviewed the metadata model and provided us with feedback. The following items are the general results of the survey:

- The metadata model describes mutually exclusive items;
- The respondents considered most items to be clearly defined and relevant. In case of confusion or ambiguity items were redefined.
- A few items were added to the model based on feedback of the respondents.
- Initially "data set" was used to describe items instead of resource. Since the model can be used to describe something else, for example a set of images or an entire project, the description "data set" is replaced by "resource".

See "ABC_metadata_model_v1.0_20230331.pdf" for ABC metadata model version 1.0. Full results of the survey are available upon request.

Expert sessions

The results of the experts sessions are:

1. The eCRF contains too detailed information to be described by the metadata model, but it is important to maintain alignment. This alignment can be achieved via, e.g., Research manager, the tool where project-level descriptions are provided.
2. All levels of hierarchy in the design of a medical image catalogue using the XNAT data model could be mapped;
3. A FAIR description at the project level (so: entire projects, not sub project level) belonging to the domain of bioinformatics laboratories for publication purposes is possible.

Implementation project

Although the simultaneously running implementation project has not been completed, a preliminary version of the machine readable JSON format was created. All items of the metadata model were provided with a variable name, data type, dependencies and required yes/no to support

implementation. The JSON file will be published in 2023 in the ABC metadata project GitHub repository GitHub - AmsterdamUMC/ABC-metadata-project.

Final findings of the project regarding the web interface for DataverseNL and the metadata exchange between two systems (DataverseNL and myDRE) will be presented separately in 2023.

Conclusion

From the expert session we concluded that there is no granularity in the model, but none is needed. The model works for each item, regardless of the level being described. It is not a direct implementable solution to everything. However, in the concept definitions and the JSON translation there is guidance to implement the model for any future known and yet unknown platform or context for biomedical research.
It's strength is that not everything is described in detail, but it can grow to become a standard for a FAIR minimum of biomedical metadata.

The model has been tested in a very limited way and more use cases are needed for fine tuning. The model in its current form is far from complete, for example the category 'synthetic data' has to be elaborated. Also the need to describe software FAIR and concisely is still untouched.