doi:10.3233/SHTI210829

What Metadata? Defining Different Types of Digital Assets as Application Targets of Metadata in Clinical Research Informatics

Matthias LÖBE[a,1]
[a] *Institute for Medical Informatics (IMISE), University of Leipzig, Germany*

Abstract. The term 'metadata' is mentioned in every one of the FAIR principles. Metadata is without question important for findability, accessibility, and reusability, but essential for interoperability. Standardized schemas have been developed by various stakeholders for decades, but too rarely come to practical use. The reason for this is that the application domain is not clearly understood. In many bio-medical research projects, the need for metadata is recognized at some point, but there is not only a lack of overview of existing standards, but also a lack of correct assessment of what individual metadata schemas were actually made for. This paper differentiates different application scenarios for metadata in clinical research.

Keywords. Metadata, Controlled Vocabularies

1. Introduction and Method

Metadata has a rather mystical meaning for many clinical researchers. The term itself does not have an undisputed definition, often scientists talk of "data about data", which on the one hand is rather general, on the other hand excludes objects that are not data. Computer scientists value metadata as necessary artifacts to exchange data between information systems without loss of information. Especially in the field of clinical research, data collection is carried out with great human, financial and regulatory cost, and there is a growing awareness that clinical studies should be registered prospectively and that results should be published even if they fail. Furthermore, there are increasing voices calling for the leakage of primary data, including individual patient data in de-identified form, to appropriate researchers.

Many research groups therefore store data sets and documents that form the basis for publications in central research data repositories, in which access can be granularly regulated according to protection needs. However, the pure instance data are only valuable for secondary analyses if the collected medical concepts behind the variables and the structure of the conducted research can be interpreted. A variety of different standards, metadata vocabularies, and medical terminologies are available for this purpose - both applicable to generic digital assets and specific to particular subfields such as clinical trials or health care data. However, if one looks at the practical use of

[1] Corresponding Author, Matthias Löbe, Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE), Universität Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany; E-mail: matthias.loebe@imise.uni-leipzig.de.

standardized metadata schemas, for example among the 2,700 research data repositories listed under the meta registry re3data [1], only a fraction uses generic metadata schemas such as Dublin Core or DataCite. Subject-specific vocabularies are even rarer by orders of magnitude. The hypothesis of this work is that many researchers do not sufficiently consider what type of digital assets they want to describe in the first place and therefore do not really use appropriate standards, which then have to be modified and extended, ultimately limiting interoperability. Using an expert-based approach, different types of assets were identified and organized.

2. Results and Discussion

Three main groups and nine subgroups of digital assets in clinical studies can be distinguished, for which very different metadata standards are relevant:

1. Structural description of data objects (structured data or documents)
    1.1. Design of the experiment (arms, cohort definition, endpoints, study sites)
    1.2. Timing of the experiment (phases, collection events)
    1.3. Structure of data collection (data models, forms, instruments, item groups, data elements, code lists)
2. Administrative description primarily for research data management
    2.1. Projects, agents, and stakeholders
    2.2. Data sets (databases) and data distributions (snapshots)
    2.3. Information systems (portals, repositories) and the catalogs they contain
3. Annotation for data usage
    3.1. Provenance (data origin, transformations, measuring methods)
    3.2. Data quality (validation and curation)
    3.3. Availability (restrictions on reuse: legal basis, patient consent)

For all these groups, metadata schemas can be found that ensure a widely accepted semantic foundation through internationally agreed standards and medical terminologies. Several candidates exist for each group; however, explaining and classifying them is beyond the scope of this paper and is the goal of future work. However, it is important to choose a suitable standard that fits the corresponding group. Otherwise, there will quickly be a need for project-specific modifications and extensions that will hamper true interoperability. Better than overambitious in-house developments is the use of coordinated vocabularies as stated in FAIR Principle R1.3: (Meta)data meet domain-relevant community standards [2], in order to develop best practices of the application of precise metadata elements in the medium term and to keep the effort for submitters low as well as for consumers.

Acknowledgments: This work was supported by the German Research (DFG grants WI 1605/10-2) and the European Commission (H2020 grant 824666).

References

[1] re3data.org - Registry of Research Data Repositories. https://doi.org/10.17616/R3D
[2] Wilkinson MD, Dumontier M, Aalbersberg IJ, et.al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18.