1st Conference on Research Data Infrastructure

Harmonizing RDM

https://doi.org/10.52825/CoRDI.v1i.271

Published: 07 Sept. 2023

Modelling Scientific Processes with the m4i Ontology

Dorothea Iglezakis[1][https://orcid.org/0000-0002-8524-0569], Džulia Terzijska [2][https://orcid.org/0000-0002-1698-6826], Susanne Arndt[3][https://orcid.org/0000-0002-1019-9151], Sophia Leimer[4][https://orcid.org/0000-0001-6272-204X], Johanna Hickmann[5][https://orcid.org/0000-0002-7535-8344], Marc Fuhrmans[6][https://orcid.org/0000-0002-9826-018X] and Giacomo Lanza[5][https://orcid.org/0000-0002-2239-3955]

[1] Universität Stuttgart, Germany

[2] TU Braunschweig, Germany

[3] Technische Informationsbibliothek, Germany

[4] Universität Duisburg-Essen, Germany

[5] Physikalisch-Technische Bundesanstalt, Germany

[6] Technische Universität Darmstadt, Germany

Abstract. We present an approach to document research data in a human and machine readable way by creating JSON-LD metadata files based on the m4i ontology. m4i is based on top level ontologies and reuses concepts of widely accepted ontologies to embed information modelled in m4i in larger contexts like a knowledge graph connecting research data with projects, actors, methods, tools and publications. We use a real-life research example from the engineering domain to show how to describe a research process with its object of research, the different steps with input and output data, the actors, and the used methods and tools. The resulting metadata files can serve as low-threshold documentation in a file system, as an exchange format between tools, as an input for data repositories and as a source of information to be used by scripts and tools.

Keywords: Metadata, Ontology, Provenance tracking, Interoperability, JSON-LD, Documentation

1. Motivation

Metadata4Ing (m4i) is an ontology for a process-based description of research activities and their results, focusing on the provenance of both research data and material objects. In engineering and many other disciplines, there is no standard yet to describe data sets in a fully semantic way. Reasons are the lack of knowledge on how to employ ontologies in everyday research and the lack of tools supporting this process. But a lot of engineers – especially in scientific computing – are used to handle data in JSON formats. Particularly suitable for that purpose is JSON-LD [1], a format that adds semantic information to normal JSON data in form of a so called context file. Through the semantic context, metadata in JSON-LD files are machine readable and actionable thanks to the exact semantic identification of classes and properties. At the same time, the files remain readable and manageable for humans, at least for humans used to handle structured data in a JSON format. We will demonstrate this approach using a real-life research example.

2. Metadata4Ing at a Glance

m4i is intended as a general process model that allows a flexible description of research activities and their results applicable to many disciplines. As depicted in Figure 1, m4i offers a selection of unambiguous well-documented terms for general concepts like processing steps, in- and output, employed methods and tools that allows modelling information about research processes and results in a structured, consistent and machine-actionable way. All the terms are available on the NFDI4Ing Terminology Service [2].

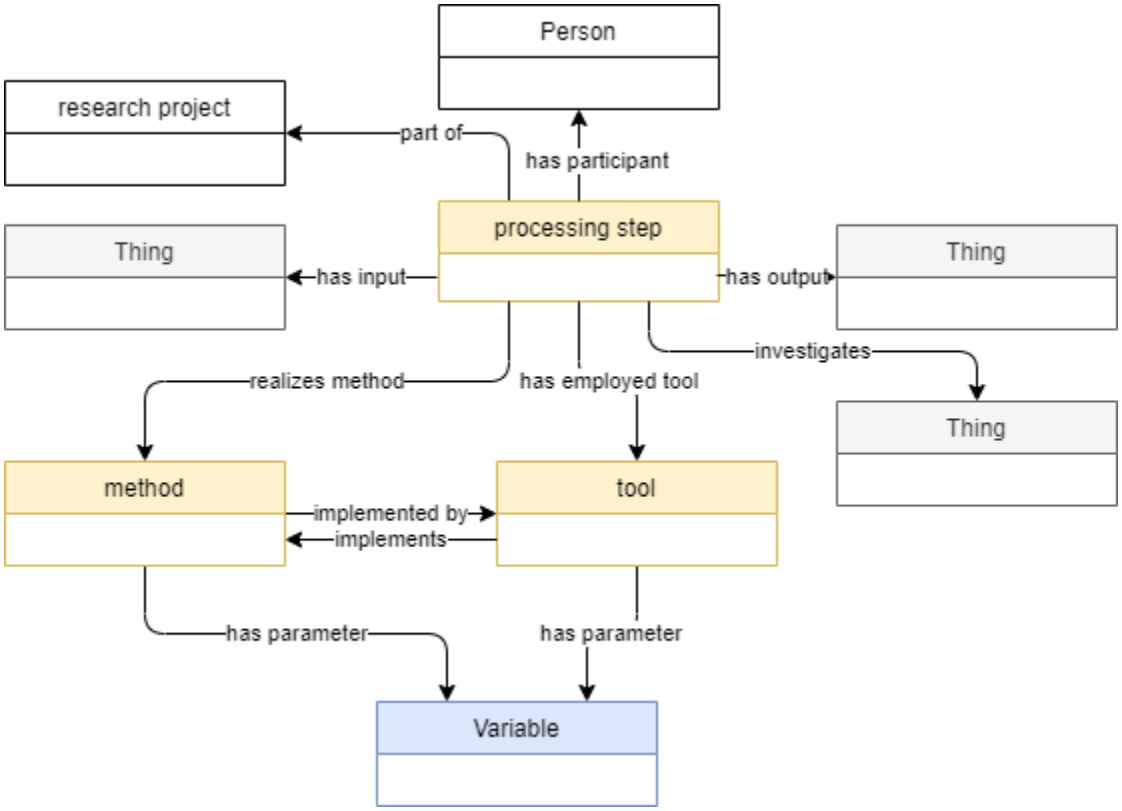

Figure 1: Core classes of m4i to describe scientific processes

One of the main benefits of using m4i is that the resulting description is highly interoperable and allows integration of data from very different scientific disciplines into a single knowledge graph. m4i has a high degree of compatibility with well-known top-level ontologies by deriving the main classes from the Basic Formal Ontology (BFO) [3], schema.org [4], the PROV Ontology (PROV-O) [5] and the Data Catalog Vocabulary (DCAT) [6], which seamlessly embeds information modelled in m4i in larger contexts. This approach is a prerequisite for FAIR (meta)data [7], especially for their interoperability.

An overview of m4i classes and properties is available at the ontology documentation [8]. The ontology code is developed at m4i's GitLab repository [9] where also its releases are published, proposals for further development in the form of issues can be made, or information how to contribute to m4i can be found.

3. Modelling scientific processes with m4i

In the talk, we will use an example of a material examination by micro X-ray computed tomography, to show how a research process can be modelled with m4i resulting in a JSON-LD metadata file.

3.1. Example setting

In the example research process, a sample of a material (in this case asphalt concrete) is examined with an XRCT scanner. The whole process consists of four steps:

1. preparation and positioning of the sample, and configuration of the parameters,
2. data generation in form of the XRCT scan,
3. image processing with the help of reconstruction algorithms and
4. post processing of the data.

The experimental setup, consisting of a holder for precise positioning of the sample, an X-ray source and a detector, is described in detail by Ruf and Steeb [10]. A resulting dataset can be found in [11].

3.2. Easy application via human-readable metadata

We will show how to create a machine-readable JSON-LD file that documents the example research process with its object of research, the different steps with input and output data, the actors, and the used methods and tools. A central element is the use of a context file, which gives the metadata the semantic context that clearly defines what exactly is meant by the information in the file, but allows to refer to classes, properties or instances via their human-readable labels. That way, semantic modelling of scientific processes without deeper knowledge in ontologies and RDF becomes possible.

The resulting metadata files can serve, for example, as low-threshold documentation / description in a file system, as an exchange format between electronic lab books and data sets, as an input for data repositories and as a source of information to be used by scripts and tools.

This information is also available in a first steps guide [12] demonstrating this approach for people with an IT affinity, e.g. application developers, research software engineers, data stewards, and tech-savvy domain experts.

4. Conclusion

Advantages of documenting research data with m4i comprise that it contributes to the implementation of good scientific practice [13], makes use of consistent metadata when searching for, analysing or otherwise using the data, and benefits collaborative work. RDF metadata can be stored as JSON-LD [1], together with the research data or code it describes. This format offers semantically enriched information that is understandable by humans and machines. In addition, a machine-actionable documentation of the data also facilitates publishing or archiving data in data repositories in a citable way. Depending on the functionalities of a data repository, the metadata contained can be used to automatically create or update a dataset. Another benefit of semantic enrichment of data sets arises when included in a global knowledge graph. The structure of m4i allows to describe and connect datasets, persons, projects, methods and tools and therefore supports complex search queries, improving the retrieval of information, especially as data pools continue to grow and connect. With the first steps guide, we aim both at researchers who want to describe their research data in a sustainable way that makes use in a knowledge graph possible, but also at the developers of tools that embed the documentation of research processes and results in the scientific workflow.

Data availability statement

The ontology m4i is available on the NFDI4Ing Terminology Service https://terminology.tib.eu/ts/ontologies/m4i and is published together with documentation, examples and guides on:

S. Arndt, B. Farnbacher, M. Fuhrmans, S. Hachinger, J. Hickmann, N. Hoppe, M. T. Horsch, D. Iglezakis, A. Karmacharya, G. Lanza, S. Leimer, J. Munke, Johannes, D. Terzijska, J. Theissen-Lipp, C. Wiljes, & J. Windeck (2022). "Metadata4Ing: An ontology for describing the generation of research data within a scientific activity". (1.1.0). Zenodo. https://doi.org/10.5281/zenodo.7706017

Underlying and related material

The documentation of m4i is available under:

https://w3id.org/nfdi4ing/metadata4ing/

The context file of m4i is available under:

https://w3id.org/nfdi4ing/metadata4ing/m4i_context.jsonld

The first steps guide is available under:

https://git.rwth-aachen.de/nfdi4ing/metadata4ing/metadata4ing/-/blob/1.1.0/training/first-steps-guide.md

Author contributions

All authors contributed equally to the conceptualization of the work described and on the review and editing of the abstract. Dorothea Iglezakis wrote the first draft together with Džulia Terzijska, Sophia Leimer, Marc Fuhrmans and Johanna Hickmann.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors (Metadata4Ing Workgroup) would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713.

Acknowledgement

The authors would like to thank Silvio Peroni for developing LODE, a Live OWL Documentation Environment, Daniel Garijo for developing Widoco and the Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine for developing Protégé.

References

1. M. Sporny , D. Longley, G. Kellogg, M. Lanthaler, P.-A. Champin, N. Lindström (2020). "JSON-LD 1.1. A JSON-based Serialization for Linked Data". https://www.w3.org/TR/json-ld11/ (accessed 2023-04-19).
2. “NFDI4Ing Terminology Service”. https://terminology.nfdi4ing.de/ts/ (accessed 2023-04-20)
3. B. Smith, P. Grenon, A. Ruttenberg, M. Brochhausen, W. Ceusters, M. Courtot, R. Dipert, J. Hastings, C. Mungall, F. Neuhaus, D. Natale, N. Otte, J. Overton, B. Peters, R. Rudnicki, S. Schulz, S. Seppälä, H. Stenzhorn, J. Zheng (2020). "BFO Basic Formal Ontology". https://basic-formal-ontology.org/ (accessed 2023-04-19).
4. “Schema.org”. https://schema.org/ (accessed 2023-04-19).
5. T. Lebo, S. Sahoo, D. McGuinness. “PROV-O: The PROV Ontology”. W3C Recommendation 30 April 2013. https://www.w3.org/TR/prov-o/ (accessed 2023-04-19).
6. D. Browning. “DCAT 2 Vocabulary”. W3C. https://www.w3.org/ns/dcat (accessed 2023-04-19).
7. M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A. ’t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson; J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, 'The FAIR Guiding Principles for scientific data management and stewardship', Scientific Data, vol. 3, pp: 160018. https://doi.org/10.1038/sdata.2016.18
8. Metadata4Ing Workgroup. “Metadata4Ing: An ontology for describing the generation of research data within a scientific activity”. Revision 1.1.0. https://metadata4ing.org/ (accessed 2023-04-19).
9. “metadata4ing”. GitLab RWTH Aachen. https://git.rwth-aachen.de/nfdi4ing/metadata4ing/metadata4ing (accessed 2023-04-19).
10. M. Ruf, & H. Steeb. "An open, modular, and flexible micro X-ray computed tomography system for research". Review of Scientific Instruments, vol. 91 no.11, 2020, p. 113102. https://doi.org/10.1063/5.0019541
11. M. Ruf, & H. Steeb. "micro-XRCT data set of open-pored asphalt concrete." DaRUS. 2020. https://doi.org/10.18419/darus-639
12. “How to Use Metadata4Ing - First Steps Tutorial”. https://git.rwth-aachen.de/nfdi4ing/metadata4ing/metadata4ing/-/blob/1.1.0/training/first-steps-guide.md (accessed 2023-04-20)
13. Deutsche Forschungsgemeinschaft. Guidelines for Safeguarding Good Research Practice. Code of Conduct. 2022. https://doi.org/10.5281/zenodo.6472827