1st Conference on Research Data Infrastructure

Harmonizing RDM

https://doi.org/10.52825/CoRDI.v1i.279

Published: 07 Sept. 2023

Machine-Actionable Metadata for Software and Software Management Plans for NFDI

Olga Giraldo[1][https://orcid.org/0000-0003-2978-8922], Danilo Dessi[2][https://orcid.org/0000-0003-3843-3285], Stefan Dietze[2][https://orcid.org/0009-0001-4364-9243], Dietrich Rebholz-Schuhmann[1,3][https://orcid.org/0000-0002-1018-0370], and Leyla Jael Castro[1][https://orcid.org/0000-0003-3986-0510]

[1] ZB MED Information Centre for Life Sciences, Cologne, Germany

[2] GESIS Leibniz Institute for Social Sciences, Cologne, Germany

[3] University of Cologne, Cologne, Germany

Abstract. Research data is on its way to be recognized as a first-class citizen in research; however, and despite its importance for science, software still has a long way to go. Recent initiatives are paving the way, including FAIR for Research Software and Software Management Plans. A step further towards machine-actionability is adding a structured metadata layer. Here we discuss some metadata elements useful to represent software and integrate it into management plans, and how it could be of benefit for NFDI.

Keywords: Research Software, Management Plan, Metadata, Machine-Actionable

1. Background

Traditionally, research outcomes have been published in text-based scholarly publications, where data and software used (or produced) are (sometimes) briefly discussed. Rich metadata exists for scholarly publication, making it easier to extract data and use it to create insights and knowledge out of it, for instance co-citation or co-author networks. Combined with Natural Language Processing techniques, in particular text-mining and text-based embeddings, further analysis becomes possible. Data and software are nowadays recognized as key players for the advance of science; however, they are not yet first-class citizens when it comes to publication and citation.

The Findable, Accessible, Interoperable and Reusable (FAIR) guiding principles for data [1] favor the use of machine-actionable metadata, i.e., metadata semantically structured facilitating search and retrieval while also facilitating (semi)automatic integration and validation. FAIR principles have also boosted research data publication and citation. Although lagging behind, research software is also moving forward in this direction, one of the reasons being its importance in science reproducibility. Some efforts working to make software a first-class citizen in research are the community-endorsed FAIR principles for Research Software [2] released in 2022, initiatives for Software Management Plans (SMPs) [3, 4] and machine-actionable SMPs [5], and best practices [6, 7] or efforts to automatically mine software citations and create structured machine-interpretable knowledge about software usage [8]. The importance of research software is also recognized in the National Research Data Infrastructure (NFDI) in Germany with groups such as the NFDI-Research Software Engineer (NFDI-RSE).

2. Metadata for Research Software

Structured, semantic, and machine-actionable metadata is a must when it comes to the FAIR principles, either for data or software. Metadata makes it easier for aggregators, archives and registries to provide a quick and open overview event if the described object is not openly available. Metadata for research software should be simple and flexible, focusing on those common elements, across software produced in a variety of scientific disciplines. A good starting point are the FAIR principles as they already suggest some metadata elements such as identifier, license, provenance and meaningful links to related objects (e.g., data consumed and produced by a software); some additional elements can be taken from scholarly publications metadata (e.g., creators, contributors, keywords).

In recent years, different efforts have repurposed Schema.org [9] for its use in science. It is a vocabulary developed by a community involving major search engines, and offers a simple way for web pages to semantically describe their content by embedding structured markup. Bioschemas [10] and CodeMeta [11] build on top of Schema.org and provide specifications to describe research software. Bioschemas ComputationalTool specification is used in bio.tools [12], while CodeMeta is used in the Software Heritage Foundation Archive [13]. NFDI-RSE is currently working on a software common marketplace that would benefit from the use of community-agreed metadata as it would enable information retrieval for software across different disciplines. A common metadata layer for research software would also make it easier to interoperate with extended and richer versions used along the consortia. For instance, the NFDI4DataScience requires additional metadata to describe training and optimization processes done with software created and/or used in solutions using data science and artificial intelligence technologies.

3. Software Management Plans

Data Management Plans (DMPs) are text-based documents describing the data management lifecycle from collection to preservation. Machine-actionable DMPs (ma-DMPs) [14] add a structured layer on top, so it becomes easier to automate the integration of information and updates. Software Management Plans pursue the same aim but for software. For instance, the Software Best Practices Focus Group, part of the ELIXIR Tools Platform, proposed an SMP for Life Sciences [3]. Similarly, the Netherlands eScience Center and the Dutch Research Council (NWO) have developed (national) guidelines for domain-agnostic SMPs [4]. Same as DMPs, SMP templates commonly pose questions to ensure that researchers follow some minimum software management standards and policies when developing research software. SMP would also help in better understanding inner workings of software, thus providing ground for a better explanation of research outcomes.

To improve interoperability and reusability of SMPs, a machine-actionable version of the ELIXIR SMP is under development [5, 15]. This ma-SMP version builds on top of the ma-DMPs so they can be easily integrated with each other. It reuses and harmonizes elements from the ma-DMP, Schema.org, Bioschemas and CodeMeta specifications, while also adding new types and properties. An overview is shown in Figure 1. In terms of NFDI, machine-actionability for DMPs and SMPs would make it easier to connect them to each other, while also developing templates tailored to different communities with a common ground, making it easier to, for instance, compare plans across different consortia and disciplines.

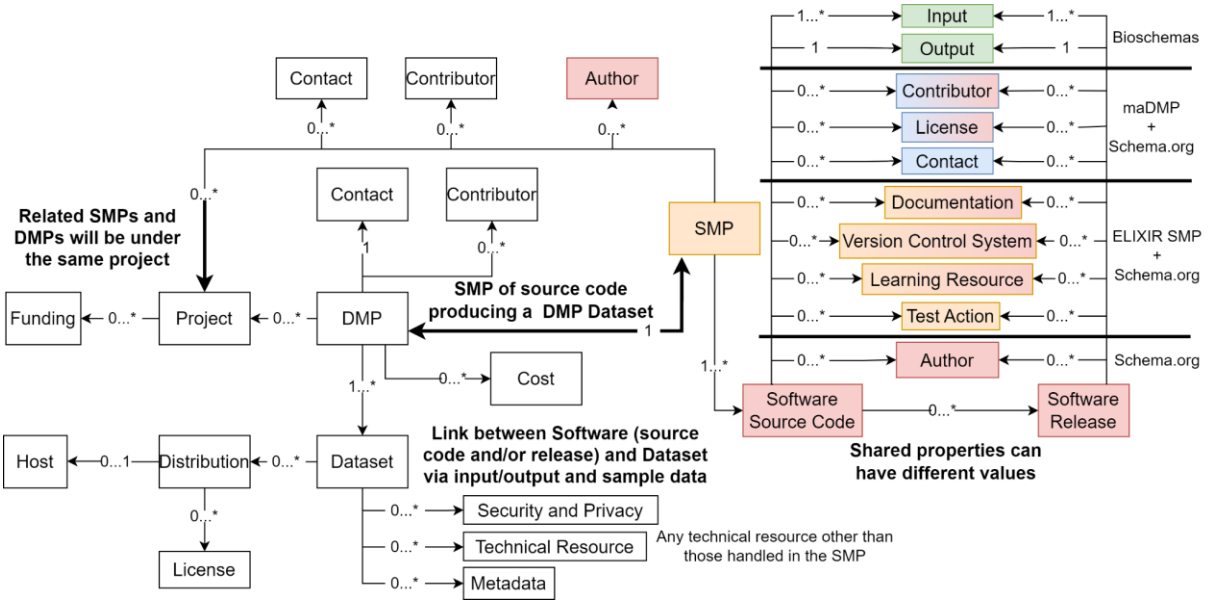

Figure 1. Metadata model for maSMP. Boxes with colored backgrounds correspond to the elements added for the maSMP case.

4. Future Work

Machine-actionability for DMPs should be embraced by the DMP working group part of NFDI-Infra section. Efforts should be combined with the RSE working group to also include maSMPs. In addition, NFDI consortia are working on the extraction of machine-interpretable metadata about software and their use in research, e.g. aiming at creating structured knowledge graphs of software and their scholarly adoption and use [8, 16]. While advancing the quality of information extraction baselines for such tasks is crucial to improve metadata quality, shared tasks on software mention detection and metadata extraction are currently being organized by the NFDI community. Using SchemaOrg as a lightweight gluing point seems reasonable as there are already efforts in that direction, it is domain-agnostic and can be customized following, for instance, the profile-way proposed by Bioschemas. By bringing multiple disciplines and communities together, NFDI is in a unique position to get community-based agreements wrt metadata for science.

Data availability statement

The metadata model corresponding to machine-actionable Software Management Plans has been published as an ontology and can be accessed at https://doi.org/10.5281/zenodo.7806639.

Author contributions

OG: conceptualization, project administration, writing – review & editing. DD: writing – review & editing. SD: writing – review & editing. DRS: conceptualization, writing – review & editing. LJC: conceptualization, funding acquisition, project administration, writing – original draft, writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Funding

The machine-actional SMP project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017536 and is part of the Research Data Alliance and European Open Science Cloud Future call 2022. NFDI4DataScience consortium is funded by the Deutsche Forschungsgemeinschaft DFG, project no.460234259.

References

1. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018. https://doi.org/10.1038/sdata.2016.18
2. Chue Hong, NP. et al. FAIR Principles for Research Software (FAIR4RS Principles). Research Data Alliance. 2022 https://doi.org/10.15497/RDA00068
3. Alves, R., Bampalikis, D., Castro, L., Fernández, J. M., Harrow, J., Kuzak, M., … Via, A. (2021, October 25). ELIXIR Software Management Plan for Life Sciences. https://doi.org/10.37044/osf.io/k8znb
4. Martinez-Ortiz C, Martinez Lavanchy P, Sesink L, Olivier BG, Meakin J, de Jong M, et al. Practical guide to Software Management Plans. Zenodo; 2022 Oct. https://doi.org/10.5281/zenodo.7248877
5. Giraldo O, Alves R, Bampalikis D, Fernández González JM, Martin Del Pico E, Psomopoulos F, et al. A metadata analysis for machine-actionable Software Mng Plans - Poster. ZB MED - Informationszentrum Lebenswissenschaften; 2023. Available: https://doi.org/10.4126/FRL01-006440396
6. Scheliga KS, Pampel H, Konrad U, Fritzsch B, Schlauch T, Nolden M, et al. Dealing with research software: Recommendations for best practices. 2019. https://doi.org/10.2312/os.helmholtz.003
7. Jiménez RC, Kuzak M, Alhamdoosh M, Barker M, Batut B, Borg M, et al. Four simple recommendations to encourage best practices in research software. F1000Res. 2017;6: 876. https://doi.org/10.12688/f1000research.11407.1
8. Schindler D, Bensmann F, Dietze S, Krüger F. The role of software in science: a knowledge graph-based analysis of software mentions in PubMed Central. PeerJ Computer Science 8:e835. 2022. https://doi.org/10.7717/peerj-cs.835
9. Guha RV, Brickley D, Macbeth S (2016) Schema.org. Communications of the ACM 59 (2): 44- 51. https://doi.org/10.1145/2844544
10. Gray AJG, Goble C, Jimenez RC (2017) From Potato Salad to Protein Annotation. ISWC Posters and Demo session. URL: http://ceur-ws.org/Vol-1963/paper579.pdf
11. Boettiger C. et al. CodeMeta: Minimal metadata schemas for science software and code, in JSON-LD. (Version 2.0). Available at https://github.com/codemeta/codemeta
12. Ison, J. et al. (2015). Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Research. https://doi.org/10.1093/nar/gkv1116
13. Abramatic J-F, Di Cosmo R, Zacchiroli S. Building the universal archive of source code. Commun ACM. 2018;61: 29–31. https://dl.acm.org/doi/10.1145/3183558
14. Miksa T, Oblasser S, Rauber A. Automating Research Data Management Using Machine-Actionable Data Management Plans. ACM Trans Manage Inf Syst. 2021;13: 18:1-18:22. https://doi.org/10.1145/3490396
15. Giraldo O., Geist L., Quiñones N., Solanki D., Rebholz-Schuhmann D., and Castro LJ. Machine-actionable Software Management Plan Ontology (maSMP ontology) (Version 0.0.1) [Dataset]. https://doi.org/10.5281/zenodo.7806639
16. Schindler, D., Bensmann, F., Dietze, S., Krüger, F., SoMeSci—A 5 Star Open Data Gold Standard Knowledge Graph of Software Mentions in Scientific Articles, 30th ACM International Conference on Information & Knowledge Management (CIKM2021), ACM 2021. https://doi.org/10.1145/3459637.3482017