Software metadata for BASE and NFDI

Authors:
Leyla Jael Castro[1], Dhwani Solanki[1]*, Nelson Quiñones[1], Lukas Geist[1], and Dietrich Rebholz-Schuhmann[1]
*Lead presenter
[1{] {ljgarcia, solanki, quinones, geist, rebholz-schuhmann}@zbmed.de, ZB MED Information Centre for Life Sciences

Abstract:
Structured, semantic, and machine-actionable metadata for all sorts of research artifacts is a must when it comes to the Findable, Accessible, Interoperable and Reusable (FAIR) principles [1]. Metadata makes it easier for search engines, recommenders, aggregators, archives, and registries to provide a harmonized and connected view across research disciplines and artifacts, including BASE4NFDI services. Metadata for scholarly publications is already consolidated thanks to efforts such as CrossRef and DataCite (although richer metadata is still possible) while metadata for data is gaining ground as data is already recognized as a key research artifact. The case of metadata for software is still trying to find its way and getting stronger thanks to initiatives such as FAIR for Research Software (FAIR4RS) [2], CodeMeta [3], Bioschemas [4], Software Management Plans (SMPs) (e.g., by ELIXIR [5] or Netherlands eScience Centre [6]) and machine-actionable SMPs (maSMPs) [7–9] (i.e., a semantic metadata layer to describe SMPs and the developed software, based on schema.org [10]).

Various efforts around metadata for software are currently discussed in NFDI sections and consortia, including the Metadata for Research Software Engineering Working Group (MetaRSE WG, Section Metadata), the nfdi.software BASE service, and the NFDI4DataScience consortium –in particular wrt maSMPs and their connection to Machine Learning (ML) Models. Related discussions in other communities include EOSC with the analysis of quality characteristics for research software and its alignments to FAIR4RS [11], and the Research Data Alliance FAIR4ML Interest Group as research software is used to create ML models. Additional efforts look to automatically extract metadata from software repositories (e.g., SoMEF [12]) or to find mentions of software in literature (e.g. SoMeSci [13]). Despite these efforts, there is still a need to better understand the common metadata layer for software across NFDI consortia, a task that is ongoing in the MetaRSE WG, plus a need to connect data and software metadata at management plans level, a task that will be soon started at NFDI4DataScience as this connection is key to enable Data Science Management Plans.

Supporting metadata for software across NFDI consortia and BASE services will facilitate the creation of knowledge graphs of software together with their scholarly adoption and use in data-driven approaches. We argue that schema.org can be used to get a common lightweight layer across multiple disciplines and consortia [14,15] while more specialized vocabularies including ontologies can be used for more in-depth analysis (that will become possible thanks to the existence of the lightweight layer).

Keywords: Metadata, Research Software, machine-actionability

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. doi:10.1038/sdata.2016.18
2. Barker M, Chue Hong NP, Katz DS, Lamprecht A-L, Martinez-Ortiz C, Psomopoulos F, et al. Introducing the FAIR Principles for research software. Sci Data. 2022;9: 622. doi:10.1038/s41597-022-01710-x
3. Jones MB, Boettiger C, Mayes AC, Arfon Smith, Slaughter P, Niemeyer K, et al. CodeMeta: an exchange schema for software metadata. KNB Data Repository. KNB Data Repository; 2016. doi:10.5063/SCHEMA/CODEMETA-1.0
4. Gray AJG, Goble C, Jimenez RC. From Potato Salad to Protein Annotation. ISWC Posters and Demo session. Vienna, Austria; 2017. p. 4. Available: http://ceur-ws.org/Vol-1963/paper579.pdf
5. Alves R, Bampalikis D, Castro LJ, González JMF, Harrow J, Kuzak M, et al. ELIXIR Software Management Plan for Life Sciences. BioHackrXiv; 2021. doi:10.37044/osf.io/k8znb
6. 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. doi:10.5281/zenodo.7248877
7. Castro LJ, Giraldo O, Geist L, Quiñones N, Solanki D, Rebholz-Schuhmann D. machine-actionable Software Management Plan Ontology (maSMP Ontology). Zenodo; 2024. doi:10.5281/zenodo.10582073
8. Giraldo O, Geist L, Quiñones N, Solanki D, Alves R, Bampalikis D, et al. A metadata schema for machine-actionable Software Management Plans. PUBLISSO-FRL; 2023. doi:10.4126/FRL01-006444988
9. Giraldo O, Dessi D, Dietze S, Rebholz-Schuhmann D, Castro LJ. Machine-Actionable Metadata for Software and Software Management Plans for NFDI. Proceedings of the Conference on Research Data Infrastructure. 2023. doi:10.52825/cordi.v1i.279
10. Guha RV, Brickley D, Macbeth S. Schema.org: evolution of structured data on the web. Commun ACM. 2016;59: 44–51. doi:10.1145/2844544
11. David M, Colom M, Garijo D, Castro LJ, Louvet V, Ronchieri E, et al. Ensure Software Quality. Zenodo; 2024 Feb. Available: https://zenodo.org/records/10723608
12. Mao A, Garijo D, Fakhraei S. SoMEF: A Framework for Capturing Scientific Software Metadata from its Documentation. 2019 IEEE International Conference on Big Data (Big Data). 2019. pp. 3032–3037. doi:10.1109/BigData47090.2019.9006447
13. 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 Comput Sci. 2022;8: e835. doi:10.7717/peerj-cs.835
14. Gray A, Castro LJ, Juty N, Goble C. Schema.org for Scientific Data. Artificial Intelligence for Science. WORLD SCIENTIFIC; 2022. pp. 495–514. doi:10.1142/9789811265679_0027
15. Castro LJ, Fluck J, Arend D, Lange M, Martini D, Neumann S, et al. Schema.org as a Lightweight Harmonization Approach for NFDI. Proceedings of the Conference on Research Data Infrastructure. 2023. doi:10.52825/cordi.v1i.280