Automated Metadata Extraction Compliant with Machine-actionable Software Management Plans

Dhwani Solanki [1], Suhasini Venkatesh [1], Dietrich Rebholz-Schuhmann [1], Leyla Jael Castro [1,*]

[1] ZB MED Information Centre for Life Sciences, Cologne, Germany
[*] Corresponding author: ljgarcia@zbmed.de

Research software and the FAIR for Research Software (FAIR4RS) principles are gaining more attention from research communities in different domains due to their role in the reproducibility of science. The Software Management Plans (SMPs) are a nice complement to the FAIR4RS principles and research software good practices. A machine-actionable layer providing semantically structured metadata describing the research software would make it easier for machines to process the data and would enable, for instance, the creation of Knowledge Graphs for research software metadata and related artifacts, e.g., data processed by the software. To this end, we have created the machine-actionable SMPs (maSMPs) metadata schema based on schema.org, and compatible with Bioschemas and Codemeta. To make it easier for researchers, we are also working on a tool to automatically extract such metadata from GitHub repositories. Here we introduce our approach towards maSMPs and present our preliminary work on automatic metadata extraction from GitHub API.

Keywords: Software Management Plans, machine-actionability, metadata

1 Introduction

Research software and its corresponding FAIR for Research Software (FAIR4RS) principles (Barker et al. 2022) are gaining more attention from research communities in different

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23951 (CC BY-SA 4.0).

domains as they contribute to the reproducibility of science. Two major obstacles linked to research software are: (i) lack of documentation with enough details to support its reuse by other researchers, and (ii) lack of structured metadata to support machine-actionability. Some of the commonly missing elements are key to FAIR4RS, for both humans and machines, e.g., authorship details, license, and unique identification. Some of these elements align with good practices that should be considered during the life cycle of research software.

To make it easier for researchers to adopt such good practices and improve the FAIR-ness of their software, the ELIXIR Software Best Practices group proposed a Software Management Plan (SMP, see Alves et al. 2021) questionnaire, i.e., a text-based document. SMPs are similar to Data Management Plans but for software rather than pure data, e.g., datasets. SMPs focus on research software good practices complementing, but not replacing, project management approaches. SMPs are meant to raise awareness of good practices while also providing means to collect information aligned to such practices. They are not a software development methodology or a full software project management approach. In fact, they can be used jointly with other approaches for software development but also on their own, which is an advantage for researchers who are not necessarily software developers, but still have to deal with code to process and data as needed for their research-oriented analyses.

While text-based SMPs serve humans well, a semantic layer was needed to enable machine-actionability, turning SMPs into machine-actionable SMPs (maSMPs). In addition to schemas supporting the machine-actionability, there is also a need to automate the metadata extraction process, making it easier for researchers to provide good basic metadata that is already present in, for instance, information about the source code repository. Here we introduce our approach towards maSMPs (Castro et al. 2024a) and present our preliminary work on a tool to extract metadata using the GitHub API (Venkatesh et al. 2025). In the future, we will expand the tool to also integrate metadata from GitLab source code repositories, and corresponding releases (for GitHub and GitLab) deposited in Zenodo.

2 Machine-actionable Software Management Plans

We have developed a maSMP metadata schema including types and properties (Castro et al. 2024a) mapping the information collected in SMPs together with usage profiles (Castro et al. 2024b), i.e., recommendations and guidelines of use. Our metadata schema builds on top of schema.org (Guha, Brickley, and Macbeth 2016), Codemeta (Jones et al. 2016), and Bioschemas (Gray, Goble, and Jimenez 2017). The first version released mid 2022 was created on top of the ELIXIR SMP (Alves et al. 2021). New versions aim to better align with other SMPs, and are commonly developed with the help of the community during hackathons. Currently, our maSMP has been aligned with the SMP

(Martinez-Ortiz et al. 2022) proposed by the eScience Centre in the Netherlands, and the SMP (Grossmann and Franke 2023) by the Max Planck Digital Library.

In Figure 1, we present the main elements of our maSMP metadata schema. As observed in the figure, the schema considers the connection between DMPs and SMPs. We do not consider elements related to the research project or budget as those have been already discussed in the maDMPs (Miksa, Walk, and Neish 2019) proposed by the Research Data Alliance DMP Common Standards Working Group. The main elements in our schema are: SoftwareManagementPlan, SoftwareSourceCode, and SoftwareApplication (corresponding to software releases).

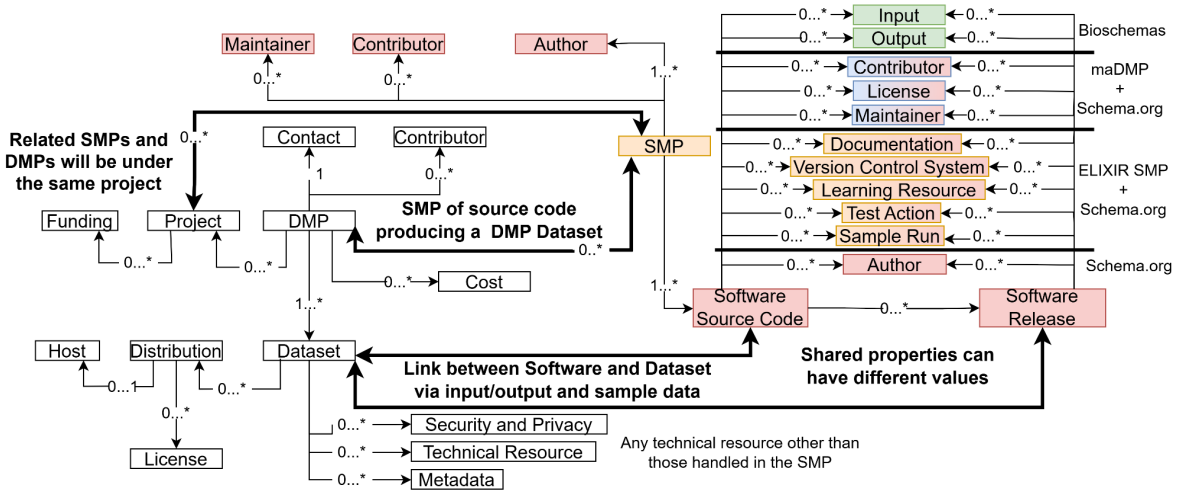

Figure 1: An extract of the maSMP metadata schema and its alignment to maDMPs.

3 Metadata extraction from GitHub

Although our maSMP metadata schema is already a step forward towards FAIR4RS and reproducibility, providing tools to automatically extract such metadata would (i) reduce the overhead for researchers and (ii) make it easier for such metadata to become part of research software repositories, e.g., hosted in GitHub, improving FAIRness of software directly at its source. There are already tools addressing the automatic extraction of metadata from GitHub repositories. For instance, the Software Metadata Extraction Framework (SOMEF, see Mao, Garijo, and Fakhraei 2019) creates a metadata file compatible with CodeMeta from information extracted via GitHub API together with machine-learning predictions, while SOMESY (Soylu et al. 2024) relies on its own file (somesy.toml) to create also a Codemeta file.

We are developing a REST API and interface (Venkatesh et al. 2025) to align metadata extraction to our maSMP metadata schema, see Figure 2. Different from SOMEF, we keep

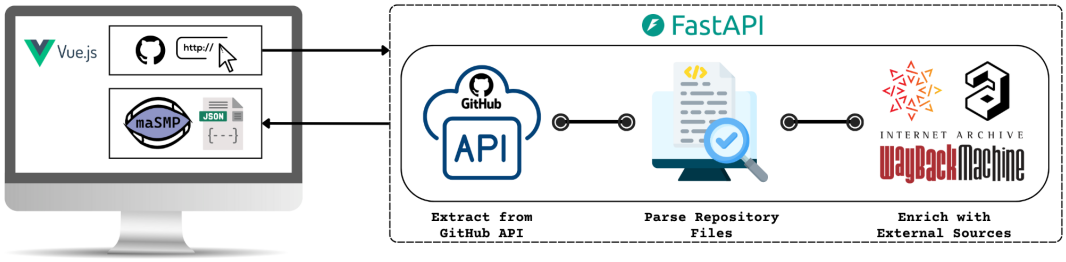

Figure 2: Our maSMP metadata extraction tool.

a clear separation between metadata obtained by direct means, e.g., GitHub API, and that one obtained as predictions using Machine Learning models. Different from SOMESY, we rely on GitHub API rather than on a toml file. At the moment, our extraction approach focuses on directly extracted metadata and distinguishes between the source code hosted in GitHub and the corresponding releases archived in external platforms. This work will help improve documentation, reproducibility, and the overall quality of research software. By automating metadata extraction and aligning it to maSMP and FAIR4RS, we aim to foster better practices in research software management across scientific domains.

4 Future work

One of the main challenges of the maSMP metadata schema is keeping it up-to-date with respect to changes in schema.org, Codemeta and SMP approaches. We will continue working with the community, e.g., via hackathons, to get feedback and improve the schema. As for the metadata extraction, the main challenge relates to the scarce data that can be directly retrieved from the GitHub API. We plan to use Large Language Models to fill in the blanks. The metadata extractor tool should also provide feedback and recommendations on good practices. Furthermore, we will work on automatic synchronization, GitLab coverage, and integration to the Research Data Management Organizer (RDMO, see Klar et al. 2025). Thanks to the integration in RDMO, together with automated metadata extraction tools, we aim to facilitate a broader adoption of maSMPs. The integration to RDMO will also enable capturing some metadata elements that otherwise would be difficult to achieve, e.g., links to tutorials about the software.

Acknowledgments

This work is part of the NFDI4DataScience project funded by the German Research Foundation (DFG), project number 460234259.

Authorship Contributions

DS, SV, DRS, and LJC contributed to the writing of the original draft, review, and editing of this manuscript. DRS contributed to the methodology. LJC contributed to the methodology and conceptualization.

Conflict of Interest

None declared.

Bibliography

Alves, Renato, Dimitrios Bampalikis, Leyla Jael Castro, José María Fernández, Jennifer Harrow, Mateusz Kuzak, Eva Martin, Fotis E. Psomopoulos, and Allegra Via. 2021. „ELIXIR Software Management Plan for Life Sciences“, https://doi.org/10.37044/osf.io/k8znb.

Barker, Michelle, Neil P. Chue Hong, Daniel S. Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, et al. 2022. „Introducing the FAIR Principles for research software“. *Scientific Data* 9 (1). https://doi.org/10.1038/s41597-022-01710-x.

Castro, Leyla Jael, Olga Giraldo, Lukas Geist, Nelson Quiñones, Dhwani Solanki, and Dietrich Rebholz-Schuhmann. 2024a. *machine-actionable Software Management Plan Ontology (maSMP Ontology).* Zenodo. https://doi.org/10.5281/ZENODO.10582073.

———. 2024b. *Usage guidance (aka profiles) for the machine-actionable Software Management Plan Ontology.* Zenodo. https://doi.org/10.5281/ZENODO.10582121.

Gray, Alasdair J. G., Carole Goble, and Rafael C. Jimenez. 2017. *From Potato Salad to Protein Annotation.* ISWC Posters and Demo Session. Visited on May 28, 2025. http://ceur-ws.org/Vol-1963/paper579.pdf.

Grossmann, Yves Vincent, and Michael Franke. 2023. „Sustainable Research Software Through Software Management Plans“, https://doi.org/10.17617/2.3525108.

Guha, Ramanathan. V., Dan Brickley, and Steve Macbeth. 2016. „Schema.org: evolution of structured data on the web“. *Communications of the ACM* 59 (2): 44–51. https://doi.org/10.1145/2844544.

Jones, Matthew Bentley, Carl Boettiger, Abby Cabunoc Mayes, Arfon Smith, Peter Slaughter, Kyle Niemeyer, Yolanda Gil, et al. 2016. *CodeMeta: an exchange schema for software metadata. KNB Data Repository.* https://doi.org/10.5063/SCHEMA/CODEMETA-1.0.

Klar, Jochen, Olaf Michaelis, David Wallace, Max Schröder, Heinz-Alexander Fütterer, Claudia Malzer, Giacomo Lanza, David Martínez Muñoz, Dario Pilori, and Harry Enke. 2025. *Research Data Management Organiser (RDMO).* Zenodo. https://doi.org/10.5281/ZENODO.596581.

Mao, Allen, Daniel Garijo, and Shobeir Fakhraei. 2019. „SoMEF: A Framework for Capturing Scientific Software Metadata from its Documentation“. In *2019 IEEE International Conference on Big Data (Big Data),* 3032–3037. IEEE. https://doi.org/10.1109/bigdata47090.2019.9006447.

Martinez-Ortiz, Carlos, Paula Martinez Lavanchy, Laurents Sesink, Brett G. Olivier, James Meakin, Maaike de Jong, and Maria Cruz. 2022. *Practical guide to Software Management Plans.* Zenodo. https://doi.org/10.5281/ZENODO.7248877.

Miksa, Tomasz, Paul Walk, and Peter Neish. 2019. *RDA DMP Common Standard for Machine-actionable Data Management Plans.* https://doi.org/10.15497/RDA00039.

Soylu, Mustafa, Anton Pirogov, Volker Hofmann, and Stefan Sandfeld. 2024. *somesy.* Zenodo. https://doi.org/10.5281/ZENODO.13120456.

Venkatesh, Suhasini, Nelson Quiñones, Dhwani Solanki, Dietrich Rebholz-Schuhmann, and Leyla Jael Castro. 2025. *v0.1.0-beta pre-release maSMPs metadata extraction.* Zenodo. https://doi.org/10.5281/ZENODO.14918500.