FAIR in practice: minimum metadata schema for bioinformatics analytics by machines

Daphne Wijnbergen[1,*], Núria Queralt-Rosinach[1], Valérie Barbié[2], Emma Verkinderen[3], Nirupama Benis[4], Annika Jacobsen[1], Peter A.C. 't Hoen[5], Claudio Carta[6], Marco Roos[1] and Eleni Mina[1]

[1]*Leiden University Medical Center, Leiden, The Netherlands*
[2]*Swiss Institute of Bioinformatics, Geneva, Switzerland*
[3]*Université Libre de Bruxelles, Brussels, Belgium*
[4]*Amsterdam UMC location University of Amsterdam, Amsterdam, The Netherlands*
[5]*Radboud University Medical Center, Nijmegen, The Netherlands*
[6]*Istituto Superiore di Sanità, Rome, Italy*

Abstract

The reuse of datasets leads to more efficient research, and a reduction of costs and time spent on generating new data. Findability and reuse of datasets as well as accessibility and interoperability can be improved by following the FAIR principles which emphasize machine actionability. In practice, metadata often lacks in machine actionability due to incomplete standardised metadata and lack of ontological descriptions. In this work, we identify minimal metadata necessary for bioinformatics tools' machine actionability and propose a schema to address current limitations. The schema includes steps for identification, selection, validation, and execution. We also align the metadata of tools to the metadata of datasets to improve machine actionability.

Keywords

FAIR data, Metadata Tools, Data Schema, Machine actionability

1. Introduction

The reuse of existing datasets allows for datasets to be optimally used, and costs to be saved. Previously, the FAIR (Findable, Actionable, Interoperable and Reusable) principles have been introduced to provide a guideline to aid in reuse. The FAIR principles highlight machine actionability in particular.

In practice, metadata is still not often machine actionable, for example due to incomplete data standards and practices. In addition, large parts of the information exist only as textual descriptions instead of as ontological descriptions. Besides datasets, FAIR is also increasingly applied in tools, for example with the FAIR4RS (FAIR or Research Software) principles. Although FAIR has been investigated for both datasets and tools separately, currently popular metadata standards for FAIR data and tools have not been investigated in a connected manner.

In this work, we identified minimal metadata necessary for machine actionability for bioinformatics analysis, and use this to propose a metadata schema that addresses limitations of current metadata for tools and data.

2. Methods

We investigated existing metadata for several tools used in rare disease research. These tools were LIMMA, ORVAL, orsum, Variomes, RD-Connect, GPAP, MOGAMUN, BridgeDB, VEP, PathVision,

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
d.wijnbergen@lumc.nl (D. Wijnbergen); n.queralt_rosinach@lumc.nl (N. Queralt-Rosinach); e.mina@lumc.nl (E. Mina)
0000-0002-7449-6657 (D. Wijnbergen); 0000-0003-0169-8159 (N. Queralt-Rosinach); 0009-0006-8085-9393 (V. Barbié); 0009-0003-9805-8998 (E. Verkinderen); 0000-0002-2101-6154 (N. Benis); 0000-0003-4818-2360 (A. Jacobsen); 0000-0003-4450-3112 (Peter A.C. 't Hoen); 0000-0003-3545-198X (C. Carta); 0000-0002-8691-772X (M. Roos); 0000-0002-8972-9206 (E. Mina)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-04-22

CEUR-WS.org/Vol-4196/paper_62.pdf

mixomics, ODAMnet and Robust Rank Aggregation. The metadata for these tools were obtained from bio.tools, The Virtual Platform of the European Joint Programme Rare Diseases, Cran.R and bioconductor.

After we recorded all the metadata, we mapped all the metadata to each other. To do this, we used the crosswalk project from CodeMeta as an alignment reference. For each resulting property in the spreadsheet, we then identified whether it is necessary for machine actionable analysis.

Finally, we executed several SPARQL queries on example metadata to demonstrate their FAIRness and machine actionability Specifically, we ran queries to identify datasets for Inclusion Body Myositis, based on their metadata and compared this to searches in the NCBI Gene Expression Omnibus. Next, we constructed queries to find tools and datasets with overlapping EDAM ontology terms. Lastly, we developed queries to identify groups of datasets by exploiting the ontological hierarchy.

3. Results

We defined a schema for the machine actionability of tools with minimal metadata. The schema is split into several steps, namely identification, selection, validation and execution. The metadata in the identification step is used for basic identification of datasets, for example with their name and description. In the selection step, metadata is used to select tools that match with the metadata or requirements of particular datasets, for example, using the category or type of the tool. In the validation step, the selection of tools can be further reduced by looking at quality indicators such as maturity and development status. Finally, metadata in the category step, such as the inputs, outputs, and operating system are used to execute the tool.

We also identified metadata that can be aligned for tools and datasets, in order to increase their interoperability, and allow for matching of datasets to tools and vice-versa. Specifically, metadata such as inputs, output, licenses and memory requirements were aligned.

Finally, to demonstrate the machine actionability of the proposed minimal metadata, we executed several SPARQL queries. In the first query, when looking for transcriptomics datasets, we found less false positives when searching for datasets using ontological terms for diseases than when searching using free text. Further, in a second query, we were able to find datasets that match the “input” metadata for tools. In the final query, we could make use of the hierarchies in ontologies to find datasets for a whole group of related diseases.

4. Conclusion

Our minimal metadata schema describes metadata necessary for machine actionability in bioinformatics for both data and tools. Additionally, we aligned the metadata for datasets and for tools, in order to improve their interoperability. We demonstrated the resulting schema with SPARQL queries.

Acknowledgments

This work was funded by ELIXIR, the research infrastructure for life science data.

Declaration on Generative AI

The authors have not employed any Generative AI tools.