OR2025 Lightning Talk (or 24x7) Track Submission

The 20th International Conference on Open Repositories, June 15-18 2025, Chicago, Illinois, USA

How do you describe software in record metadata?

Matteo Cancellieri, CORE, The Open University, matteo.cancellieri@open.ac.uk ORCID: 0000-0002-9558-9772

Petr Knoth, CORE, The Open University, petr.knoth@open.ac.uk ORCID: 0000-0003-1161-7359

Abstract

The discoverability, attribution, and reusability of open research software are often hindered by its inadequate representation in research manuscripts. Frequently mentioned only implicitly or buried within supplementary materials, software fails to achieve recognition as a distinct, citable output. Addressing this challenge requires systematic identification and assignment of persistent identifiers (PIDs) to software, ensuring compliance with FAIR (Findable, Accessible, Interoperable, and Reusable) principles. Despite its significance, most open research software remains underrepresented in metadata, with limited explicit links between software and the research papers introducing or using them.

The SoFAIR project (2024–2025) seeks to enhance the identification and representation of software assets in research. By leveraging the global network of open repositories, the project aims to look into the current state of metadata standards and proposes adaptations to include software descriptions.

The presentation will explore current metadata formats and propose actionable solutions for improving the discoverability and reusability of open research software, aligning with best practices for metadata interoperability.

Keywords

Research Software Metadata

Persistent Identifiers (PIDs)

FAIR Principles

Discoverability

Audience

The audience for this contribution includes repository managers, repository software developers and metadata experts.

Proposal

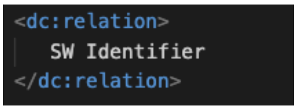
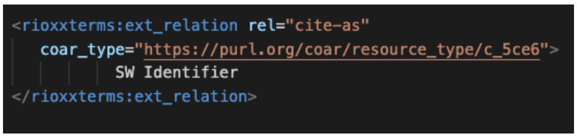
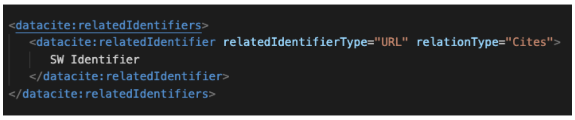
One of the primary challenges affecting the discoverability, attribution, and reusability of open research software lies in its frequent obscurity within research manuscripts. These valuable resources are often mentioned in passing, mentioned implicitly or buried in supplementary material, preventing them from being recognised as distinct, citable outputs. For research software to achieve the status of first-class bibliographic records, it should first be systematically identified and assigned persistent identifiers (PIDs), ensuring alignment with FAIR (Findable, Accessible, Interoperable, and Reusable) principles. Despite their importance, most open research software assets fall short of these principles, and explicit links between software resources and the papers introducing or utilising them remain rare and even when available, these relationships are hardly ever exposed in the metadata.

The SoFAIR project [1], running from 2024 to 2025, addresses challenges related to the identification and representation of software assets in metadata records of research manuscripts by leveraging the extensive content available across the global network of open repositories. This work focuses on presenting the current state of the art in research software representation within the metadata of research manuscripts and recommends solutions for incorporating software links into OAI-PMH metadata.

At present, no established metadata standard for research outputs explicitly supports software descriptions in its guidelines. However, existing guidelines can be adapted to include software descriptions while maintaining compliance. Failure to properly represent software could result in a loss of critical information, making it difficult for machines to differentiate software from generic links.

To ensure accurate software representation, the primary requirement is the inclusion of a URL, ideally persistent. The more accurate description of a software is the Software Heritage Identifier (SWHID), however, the SWHID tends to represent a commit in the versioning system and sometimes finding the SWHID for all the software relations might prove complex. DOIs, hdl.handles and ARK identifiers are acceptable but require the registration of the software in an extra registry. Alternatively, origin links from well-known code forges provide a practical and widely achievable solution for representing software in metadata.

The presentation will examine the current state-of-the-art metadata formats and offer recommendations for identifying research software in a machine-readable and interoperable manner. The discussion will focus on the following metadata profiles:

Metadata format	Example
Dublin Core[2]: Software can be described using "dc:relation," but the lack of semantic richness limits its utility for precise software representation.	 <dc:relation> SW Identifier </dc:relation>
Rioxx Version 3[3]: This format provides a means to represent software through the "rioxxterms:ext_relation" field and the “coar_type” attribute that allows to specify that the relation is a “software” using the COAR Resource Type Vocabulary [5]. It is currently the most precise approach for describing software, even though it wasn’t intentionally designed with software in mind and might lack semantics.	 <rioxxterms:ext_relation rel="cite-as" coar_type="https://purl.org/coar/resource_type/c_5ce6"> SW Identifier </rioxxterms:ext_relation>
OpenAIRE Guidelines[4]: While it includes provisions for software representation as an external link, it poses challenges in disambiguating software links from other related links, such as datasets or supplemental materials, especially for machine interpretation.	 <datacite:relatedIdentifiers> <datacite:relatedIdentifier relatedIdentifierType="URL" relationType="Cites"> SW Identifier </datacite:relatedIdentifier> </datacite:relatedIdentifiers>

The presentation will evaluate the strengths and limitations of each solution, to gather feedback and collaboratively design a set of recommendations for effectively representing software in research metadata records.

References

[1] https://sofair.org/

[2] https://www.dublincore.org/specifications/dublin-core/relation-element/

[3] https://rioxx.net/profiles/#rioxxterms:ext_relation

[4] https://guidelines.openaire.eu/en/latest/data/field_relatedidentifier.html

[5] https://vocabularies.coar-repositories.org/resource_types/