

Publish metadata schema used (from [July 11, 2019](#) late call)

Attendees: Alejandra, Daniel (Garijo) [<https://orcid.org/0000-0003-0454-7145>], Wesley, Stephan Druskat (stephan.druskat@dlr.de), Alain

This should contain:

1. Statement such as “we use this schema” (codemeta or schema.org or whatever) and a link to that schema’s site if it’s a standard schema or its documentation if it’s not.
 - a. ALL Versions + Version URI + documentation
2. Characteristics of the expected metadata (e.g., does a link to the license go in the field, or the name of the license) - metadata validation / datatypes
 - a. Clearly define what is required vs what is optional
 - b. Define conventions in, e.g., an FAQ (or part of the spec), such as:
 - i. Language used for annotations
 - ii. Datatype of values (ORCID ids for authors or Names)
 - iii. Number of letters in country codes
 - iv. Representation of unknowns, e.g., known unknowns (what are things a year field can be if not four numbers? ‘Unknown’, ‘n.d.’, null...)
3. The schema supports structure data that is machine-readable and may be used to validate instances (i.e., metadata instances can be validated against the schema)
4. The schema should be cross-walkable with CodeMeta (e.g., to make sure that registries can talk between themselves)

Why are we recommending this as a best practice? (why should repos have this)

Why should you state which schema you use and link to it (and its documentation)?

- So users have guidelines on how to fill in their entries (1)
- So users know what each metadata field means (1)
- So that submitting users understand that it’s important that they provide metadata in order to make their software more accessible (1)
- So that users can figure out what metadata source they can provide (e.g., users provide a CITATION.cff with their source code, which the registry can convert to CodeMeta if that’s used by the registry; or, users provide a codemeta.json they have created themselves (from CFF or in another way), and registries can validate) (1)

Why should you state the characteristics of the metadata you expect?

- So users know what to pay specific attention to (initially) without being overwhelmed.
- So that users know which metadata the registry can represent for their resource in the first place, and adapt the information they provide. (2)
- So you know what your metadata fields mean in 2 years, and new people working with/on your resource can easily learn what the fields mean and what to expect from them (in case the designer of the schema goes away) - documentation (2)

Why should the schema support structure data that is machine-readable and may be used to validate instances?

- You don't want invalid data in your repo! Machine validation is nice (cf. type-checking) and humans make mistakes sometimes. Should have automatic safeguards against that - so a well-defined metadata schema that lets you validate instances.

Why should your schema be cross-walkable?

- So that submitting users understand that it's important that the registry uses a specific schema (4)

1 or 2 examples that show this in the wild:

The 2 examples from the call linked above are:

Non machine-readable example:

<https://www.ands.org.au>

(https://www.ands.org.au/_data/assets/pdf_file/0004/728041/Metadata-Workinglevel.pdf)

<https://jats.nlm.nih.gov/publishing/tag-library/1.2/element/resource-id.html>

Machine-readable example:

<http://ontosoft.org/software> (part of repo: <http://ontosoft.org/>)

HTML: <http://ontosoft.org/software>

RDF+XML: <http://ontosoft.org/software> (curl -sH "Accept:application/rdf+xml" -L

<http://ontosoft.org/software>)

Do we know of more?

bio.tools BiotoolsSchema: <https://biotoolsschema.readthedocs.io/en/latest/> (documentation) and <https://github.com/bio-tools/biotoolsSchema/> (source for the schema)

- Zenodo's JSONSchemas:
 - For depositing: <https://zenodo.org/schemas/deposits/records/legacyrecord.json> (also in [developer docs](#))
 - For the published record: <https://zenodo.org/schemas/records/record-v1.0.0.json>