Metadata in reuse: harvesting, licensing, repurposing and FAIR

Data Description and Metadata - What it takes to produce a good one?
December 8, 2021

Tuomas J. Alaterä, FSD
tuomas.alatera@tuni.fi
https://orcid.org/0000-0002-3448-3448

cessda.eu

@CESSDA_Data

Human readable metadata is fine…

- …but having machine-readable metadata too is better (even though one cannot even see it!)
  - Allows harvesting
  - Allows automagical enriching
  - Allows wider discoverability
  - Allows citing
  - Allows building on top of interoperable metadata

  - Linked open data and APIs are crucial

Persistent Identifiers

- Identifiers for everything
  - Not only for resolving to the resource
  - PIDgraphs rely on persistent identifiers to build enriched and meaningful relations
    - Researcher IDs (predominantly ORCID)
    - Research Organisation / Funder IDs (ROR, ISNI, URN …)
    - Publications, articles and such (ISBN, ISSN, DOI, Handle…)
    - Research project IDs (RAiD…)
    - In addition, identifiers for any entities that has relations with the dataset

- Needed that the platform supports the use of IDs
- Doable for example in JSON-LD for basically all actors
- Should be included as text, if types or fields for relations not available

Persistent Identifiers, JSON-LD examples

```
"publisher": [
  {
    "@type": "Organization",
    "sameAs": "https://ror.org/040af2s02",
    "name": ”University of Helsinki"
   }
```

```
"name": [
      {
         "@value": "Finnish Voter Barobeter 1973",
         "@language": "en"
      },
      {
         "@value": "Puolueiden ajankohtaistutkimus 1973",
         "@language": "fi"
      }
  ],
```

```
"citation": {
          "@type": "CreativeWork",
          "creator": [{
                    "@id": "https://orcid.org/ 0000-0000-0000-0000",
                    "@type": "Person",
                    "name": "John Smith",
                    "familyName": ”Smith",
                    "givenName": "John",
                    "identifier": "https://orcid.org/0000-0000-0000-0000",
                    "email": "j.smith@somedomain.org"
                    },
```

Metadata served in different formats

- Offer the metadata for harvesting primarily via an API
  - Multiple formats can be produced from the core discipline specific metadata
  - Basic mapping e.g. to Dublin Core increases usability
  - FSD uses Kuha2 for serving DDI Codebook, EAD3 and OAI Dublin Core
    - DataCite is a recommended format to consider
- Metadata in XML or different LOD format can be embedded or linked to the landing page

Licenses for metadata

- Licenses or conditions for use for data are common
- Metadata should be licensed as well
  - Because of clarity of (re)use (and perhaps merit)
  - At times required by aggregators
  - Recommended formats CC0 and CC BY
  - Metadata license declared in machine-actionable format
  - Like data, metadata should be persistent, and versions monitored

Citation

- Merit depends on citations
- Therefore, data citations must stand the test of time
  - Use of PIDs
  - Repository driven service: offer both a citation example and a citation in machine-actionable format
  - Importance of a landing page for data where relevant information is available, like ID, title, creator, publisher, release date, version.
  - Make sure these are available in machine-readable format using e.g. schema.org or DC.

Machine-actionable metadata and FAIR

- FAIR depends on machine-actionable metadata and the use of various (FAIR) controlled vocabularies
- In SSH domain, tools offered by CESSDA (vocabularies.cessda.eu) are of use (DDI, CESSDA vocabs)
- Other vocabularies e.g. for place names, coverage, specimen etc. as needed by the community
- Expressed in a format suited to your needs (using schema.org, DC Terms, Open Graph…)

In Conclusion

- Interoperability is both a technical and content issue
- PIDs and other published identifiers need to be collected early
- Further PIDs need to be minted as needed
- Different forms of metadata may be needed for harvesting
- For machine-actionability a standard for open linked data is needed
- Machine actionable interoperability relies on various sources and should not depend on a manual processing only