Harmonizing the Implementation of PIDs across Repositories

Andrea Pörsch[1] // Kirsten Elger[1] // Emanuel Söding[2] // Dorothee Kottmeier[3] // Stanislav Malinovschii[2] // Yousef Razeghi[4] // Sören Lorenz[2]

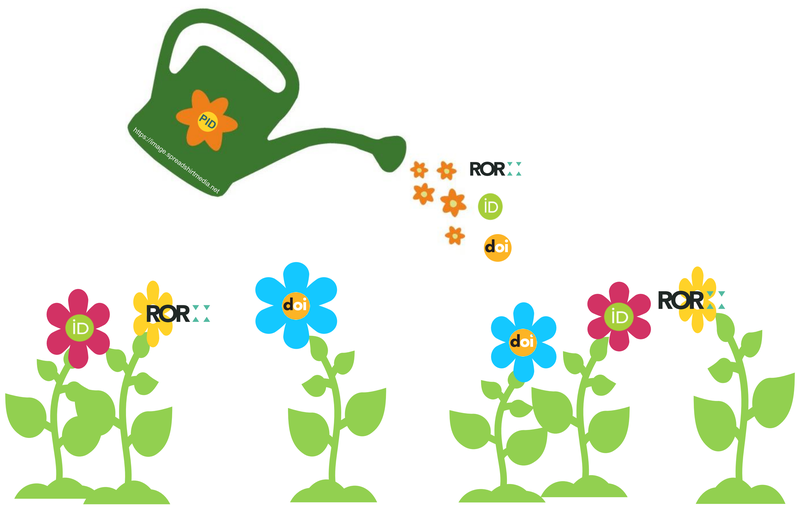
0000-0003-4502-6223 0000-0001-5140-8602 0000-0002-4467-642X 0000-0002-4263-4234 0009-0002-9792-6768 0000-0002-0007-630X 0000-0001-8577-6614

1 GFZ German Research Centre for Geosciences
2 GEOMAR Helmholtz Centre for Ocean Research Kiel
3 Helmholtz Centre for Polar and Marine Research - AWI
4 Helmholtz Centre for Environmental Research - UFZ

In our increasingly digital and interconnected world, the integration of Persistent Identifiers (PIDs) in metadata are essential for machine-readable and -understandable metadata as also described in the FAIR Guiding Principles for research data management. PIDs provide unique, permanent and machine-readable references to various types of digital objects, including publications, datasets, scientific software, individuals, organizations, samples that together represent the broad range of research outcomes.

How can we encourage the use of identifiers?

Within the AK Metadata-PIDs working group (a joint initative between the HMC Hub Earth and Environment and the Helmholtz DataHub Earth and Environment), we discussed several PID systems and reached a consensus on recommending specific systems for different purposes: "ORCID" for identifying individuals, "ROR" for organizations, and the "PIDINST" PID for instruments. Our working group has focused on supporting the ongoing PID implementation in research infrastructures by conserving existing, well-established PID implementations (best practices) and promoting their integration in future systems. We further aim to provide support and guidance for new implementations.



What are the differences between metadata schemas?

The two metadata schemas most used in the Earth and Environment sciences are DataCite and ISO 19115 (INSPIRE – EU Comm).

Both schemas are rich and comprehensive for their use cases. While ISO 19115 is dedicated to describe “classical” geodata (spatial data) and is extensively used by geological surveys and agencies, the DataCite schema was developed for DOI registration and is based in the research context. Especially the DataCite Schema has been significantly further developed to support data discovery, the use of PIDs and citations (with clear guidelines of how to include them in the metadata).

In contrast to this, ISO 19115 metadata allows for high granular description of individual data points (e.g., with the information about connected spatial and temporal information – which is not possible in the DataCite schema at the moment), but lacks, e.g. a clear option to add citations of publications or datasets to the metadata. Consequently, the ISO Metadata available via the assessed repositories are much less harmonized for these properties than the DataCite metadata files. Do we need to change this at all?

DataCite XML Schema

```
<relatedIdentifiers>
  <relatedIdentifier relatedIdentifierType="DOI" relationType="IsReferencedBy">10.1007/s001900050078</relatedIdentifier>
  <relatedIdentifier relatedIdentifierType="DOI" relationType="IsReferencedBy">10.1007/978-3-642-78149-0_49</relatedIdentifier>
  <relatedIdentifier relatedIdentifierType="DOI" relationType="IsDocumentedBy">10.2312/GFZ.b103-16087</relatedIdentifier>
  <relatedIdentifier relatedIdentifierType="DOI" relationType="References">10.5880/igets.su.l1.001</relatedIdentifier>
  <relatedIdentifier relatedIdentifierType="DOI" relationType="IsCitedBy">10.1007/s40328-018-0212-5</relatedIdentifier>
</relatedIdentifiers>
```

publications
data report
dataset

ISO 19115 XML Schema

```
<gmd:aggregationInfo>
  <gmd:MD_AggregateInformation>
    <gmd:aggregateDataSetIdentifier>
      <gmd:RS_Identifier>
        <gmd:code>
          <gco:CharacterString>10.2312/GFZ.b103-16087</gco:CharacterString>
        </gmd:code>
        <gmd:codeSpace>
          <gco:CharacterString>DOI</gco:CharacterString>
        </gmd:codeSpace>
      </gmd:RS_Identifier>
    </gmd:aggregateDataSetIdentifier>
    <gmd:associationType>
      <gmd:DS_AssociationTypeCode codeList="http://datacite.org/schema/kernel-4"
      codeListValue="IsDocumentedBy">IsDocumentedBy</gmd:DS_AssociationTypeCode>
    </gmd:associationType>
  </gmd:MD_AggregateInformation>
</gmd:aggregationInfo>
```

Identifier
Identifier type (DOI)
Relation type with reference to the DataCite schema

Related identifiers in the DataCite and ISO 19115 Metadata Schema. While DataCite provide a dedicated „relatedIdentifier“ property, in the ISO 19115 example, this information is provides a „Aggregate Information“ property. The ISO schema has the flexibility to link to external sources, here the relation type from the DataCite schema. This is one example for adding citations in the ISO metadata from GFZ Data Services (Neumeyer et al., 2017 https://doi.org/10.5880/igets.po.l1.001). Other repositories use different properties

Summary and Outlook

The metadata schemas ISO19115 and DataCite have been developed by different communities and for different purposes. Both have their strengths and weaknesses. It is possible to map a subset of metadata properties in a harmonized way (e.g. keywords, authors information), some metadata properties remain challenging (e.g. citations or PIDs in ISO metadata, dedicated spatial/temporal context in DataCite metadata). The full record, however, cannot be mapped. As a result, the content of the ISO metadata representations vary much more between the repositories than the DataCite metadata.

Do we need to align these schemas to be able to map the full metadata between them? For this, we need to further explore the use cases and especially the connections between the communities. If more geological surveys used DOIs, it would be reasonable to expand the ISO metadata schema for the inclusion of PIDs like ROR, ORCID and related references (citations). The DataCite Metadata Working group is already exploring option to be more specific and actively looks into properties already used by ISO 19115 (pers comm, K. Elger).