Towards richer metadata – perspectives from three DataCite projects

Kelly Stathis, **Xiaoli Chen**, **Cody Ross**, **Paul Vierkant**

Blog post published November 17, 2022 in *DataCite Blog - DataCite*
https://doi.org/10.5438/fz2e-fs29

Abstract

Metadata Metadata is at the heart of DOIs and open scholarly infrastructure. At DataCite, our metadata schema defines what metadata properties can be included through DOI registration. The schema currently includes just six required properties—identifier (the DOI), creator, title, publication year, publisher, and resource type—along with 14 recommended and optional properties.

Keywords

Information Systems (Subfield), Research Data Management Practices (Topic), FAIR Workflows, IGSN, Metadata, NFDI4Ing, Projects

Recommended citation

Stathis, K., Chen, X., Ross, C., & Vierkant, P. (2022, November 17). Towards richer metadata – perspectives from three DataCite projects. *DataCite Blog - DataCite*. https://doi.org/10.5438/fz2e-fs29

Copyright

Copyright © Kelly Stathis et al. 2022. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Metadata

Metadata is at the heart of DOIs and open scholarly infrastructure. At DataCite, our metadata schema defines what metadata properties can be included through DOI registration. The schema currently includes just six required properties—identifier (the DOI), creator, title, publication year, publisher, and resource type—along with 14 recommended and optional properties.

On the one hand, requiring only six metadata properties keeps the schema flexible and makes it easy to get started with DOI registration. At the same time, we want to encourage all DataCite Metadata Schema users to go beyond the mandatory properties and to share rich metadata that includes all available information about a given resource. This is especially important for metadata properties that are essential for discoverability—such as description and subject—and building connections between PIDs—including identifiers for related resources, people, and organizations. Keeping metadata up-to-date is also critical to ensure that the "persistent" part of persistent identifiers lives up to its full potential.

> DataCite Connection Metadata. An infographic shows five bubbles that represent an entity and its persistent identifier: researchers with the ORCID iD; related research outputs with a DOI; research funders with the ROR ID and Crossref Funder ID; research organizations with the ROR ID. These four bubbles are floating around and are linked to the inner bubble of research output with DataCite DOIs.

We know it isn't always easy to create, enhance, and maintain robust metadata. Among other challenges, there is a need to collaboratively define metadata best practices so that metadata creators, repository platform providers, and the open research community can create metadata with confidence. At DataCite, we are involved in several projects and partnerships that support rich metadata across different domains. This post is based on the November DataCite Open Hours, where we heard from team members involved in three projects/partnerships: the Implementing FAIR Workflows Project; the IGSN-DataCite Partnership; and the NFDI4Ing Seed Funds project.

FAIR Workflows

In the Implementing FAIR Workflows project, we look at metadata completeness from two perspectives: 1) capturing the metadata of outputs that traditionally fly under the radar by establishing new PID workflows; and 2) enriching the metadata of outputs that are already being shared by improving existing PID workflows. For the former, we focus on helping the researchers to identify interim outputs and build sharing practices around them, at the same time, supporting repository platforms and research tools in developing integration to streamline PID registration and metadata sharing and increase interoperability between technologies. For the latter, we emphasize the use of community-endorsed resources for metadata generation and maximize coverage of metadata submission to not only include core metadata, but also recommended and optional metadata in the curation process. Organizations can take action by creating comprehensive crosswalks between their local metadata and the DataCite metadata schema, and implementing workflows that capture and share connection metadata in a standardized format.

IGSN perspective

The IGSN ID is a globally unique and persistent identifier for all types of material samples from all disciplines, including sample aggregates, destroyed/discarded samples, and even sample collection sites. As with other research outputs, metadata plays a critical role in describing and connecting IGSN IDs to maximize discoverability and reuse. Under its partnership, IGSN e.V., and DataCite have established working groups to define best practices for material sample metadata in the DataCite Metadata Schema. Furthermore, we continue to work with disciplinary samples communities to better support their needs and reach consensus on metadata standardized within and among these communities.

IGSN ID metadata in the DataCite Metadata Schema can be collected and enriched throughout every stage of the samples workflow, from planning and collection through to repository ingest and publication. Principal Investigators, analysts, curators, and repository managers are encouraged to enrich sample metadata throughout the workflow processes and over time, using metadata from field-based tools, analytical systems, and local samples databases. In contrast to many research outputs, the physical nature of samples means that they are often broken down into smaller and smaller pieces. Relationship metadata for IGSN IDs is thus valuable not only for unambiguously linking samples with related datasets, publications, researchers, institutions, and external metadata, but also vital for describing the linkages between parent samples and derived children.

NFDI4Ing

The strength of the DataCite Metadata Schema is its domain-agnostic character. As a global standard for describing research outputs and resources, the schema needs to be as general as possible but at the same time as specific as necessary. DataCite's engagement in different research projects reflects our will to connect to research domains that are in need of domain specific metadata. One goal in the NFDI4Ing Seed Funds project is to map the DataCite Metadata Schema to other schemata used in the engineering sciences as well as schema.org. Beyond mapping, we seek best practices in linking our schema to domain-specific schemata in general. The second goal of NFDI4Ing is to develop user stories around the upcoming resource type for instruments, using connection and relationship metadata. With relationship metadata, researchers and research organizations can find out which instrument was used to create a dataset. To visualize these relationships, extensive PID metadata – whether general or domain-specific – must be part of the PID graph. Similar to the projects mentioned above, the importance of metadata within the NFDI4Ing Seed funds project cannot be overstated.

What's next?

These three projects and partnerships are some of the work DataCite is doing to improve DOI metadata. In the coming months, you'll hear more from us about how you can contribute to the next major version of the DataCite Metadata Schema (5.0), along with updates as we finalize our

next minor version (4.5). We look forward to working with the DataCite community to support your efforts to collect and share richer metadata.