

# DataCite

## Metadata Basics

**Helena Cousijn**

*Director of Community Engagement*

22nd of September 2022  
DataCite Member Meeting 2022



[@datacite](https://twitter.com/datacite)

# Agenda

- What is metadata and why do you need it?
- The DataCite metadata working group & schema
- What's in the metadata schema
- Openly available metadata & harvesting
- Doing more with metadata
- What's next for the metadata schema?

**What is metadata?**

data that provides information about other data

not the content of the data

metadata helps us understand the structure, nature, and context of the data

# Why metadata

Metadata provides information about the relevant entity

When a member registers a DOI, they always register metadata about the entity

This helps with discoverability, connectedness and FAIRness

# Why a schema?

The DataCite Metadata Schema is a list of core metadata properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions.

The schema provides standardization which allows users to search across metadata, and thereby increases interoperability

# **The schema & working group**

## What is the DataCite Metadata Schema?

- A metadata schema specifically for DataCite DOIs

## Who develops it?

- The [DataCite Metadata Working Group](#) and DataCite staff

## Where can I find it?

- All versions: <https://schema.datacite.org/>
- Latest version (4.4): <https://schema.datacite.org/meta/kernel-4.4/>

## When is it updated?

- About every 1-2 years



# The metadata working group

- consists of 12 DataCite members and 2 DataCite staff members
- is chaired by two representatives of member organizations
- meets once a month to work on the schema
- has an open call for nominations at the end of the calendar year
- can be reached by email or suggestions can be contributed to productboard

# Schema versions

- A new schema version is released once every 1-2 years
- Currently in use: 3.0, 3.1, 4.0, 4.1, 4.2, 4.3, and 4.4
- Schema 4.5 will be released early next year
- After that, we will look at major changes and work on schema 5.0

## Note:

- schema 5.0 will contain breaking changes
- At some point, support for schema 3 will be discontinued

# The DataCite Metadata Schema

## What is in the schema?

- The schema consists of 20 metadata **properties** (sometimes called “fields” or “elements”).
- Hierarchical structure: some properties have sub-properties.
- Some are mandatory, some recommended or optional.
- Some can be repeated.
- Some have controlled list values.

# The structure

## 20 metadata properties:

| <i>Mandatory</i>     | <i>Recommended</i> | <i>Optional</i>  |
|----------------------|--------------------|------------------|
| Identifier           | Subject            | Language         |
| Creator (with ORCID) | Contributor        | Alternate ID     |
| Title                | Date               | Size             |
| Publisher            | Related identifier | Format           |
| Publication year     | Description        | Version          |
| Resource Type        | GeoLocation        | Rights           |
|                      |                    | FundingReference |
|                      |                    | RelatedItem      |

# The structure, continued

## Lots of sub-properties

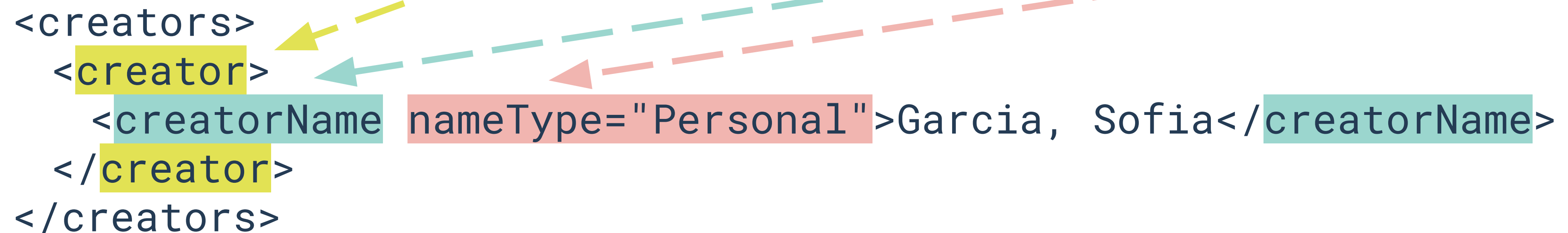
Some mandatory, some optional

Some have their own sub-properties

Some repeatable, some non-repeatable

## Metadata is primarily represented in XML

Properties are represented as **elements**; sub-properties can be **sub-elements** or **attributes**



```
<creators>  
  <creator>  
    <creatorName nameType="Personal">Garcia, Sofia</creatorName>  
  </creator>  
</creators>
```

The diagram illustrates the XML structure with annotations. A yellow box highlights the `<creator>` element, with a yellow dashed arrow pointing from the text 'elements' in the paragraph above. A teal box highlights the `<creatorName>` element, with a teal dashed arrow pointing from the text 'sub-elements' in the paragraph above. A red box highlights the `nameType="Personal"` attribute, with a red dashed arrow pointing from the text 'attributes' in the paragraph above.

## Supports a variety of resource types:

|                       |                      |
|-----------------------|----------------------|
| Audiovisual           | Model                |
| Book                  | OutputManagementPlan |
| BookChapter           | PeerReview           |
| Collection            | PhysicalObject       |
| ComputationalNotebook | Preprint             |
| ConferencePaper       | Report               |
| ConferenceProceeding  | Service              |
| DataPaper             | Software             |
| Dataset               | Sound                |
| Dissertation          | Standard             |
| Event                 | Text                 |
| Image                 | Workflow             |
| InteractiveResource   | Other                |
| Journal               |                      |
| JournalArticle        |                      |

# Beyond mandatory properties

Recommended and optional properties:

- Subject
- Contributor
- Date
- RelatedIdentifier
- Description
- GeoLocation
- Language
- AlternateIdentifier
- Size
- Format
- Version
- Rights
- FundingReference
- RelatedItem



# RelatedIdentifier

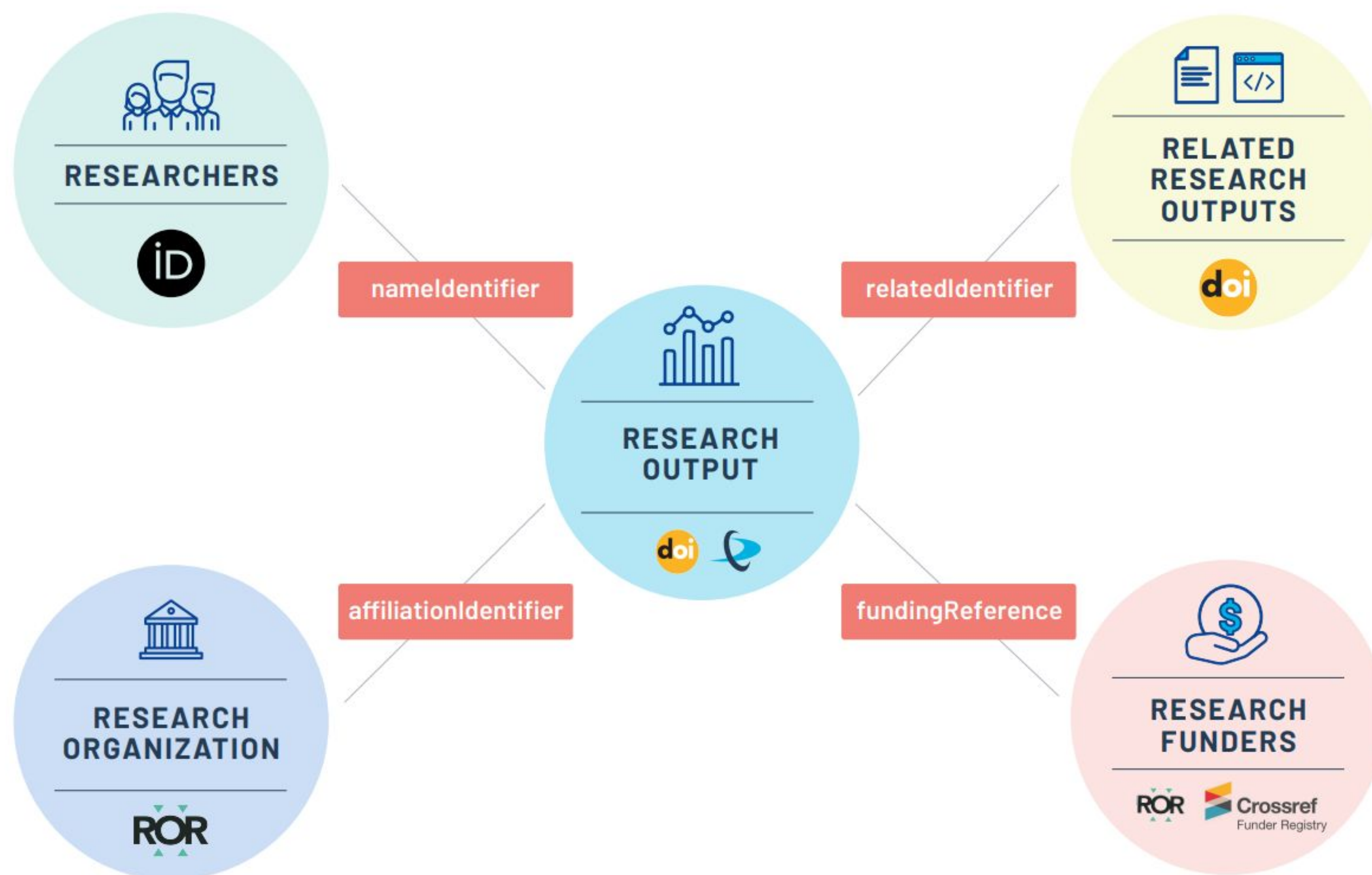
**RelatedIdentifier** is used for identifiers of related resources.

Mandatory sub-properties:

- **relationType**: the type of *relationship*
- **relatedIdentifierType**: the type of identifier

| relationTypes                        | relatedIdentifierTypes |
|--------------------------------------|------------------------|
| IsCitedBy / Cites                    | ARK                    |
| IsReferencedBy / References          | arXiv                  |
| IsSupplementTo / IsSupplementedBy    | bibcode                |
| IsNewVersionOf / IsPreviousVersionOf | DOI                    |
| HasVersion / IsVersionOf             | EAN13                  |
| IsObsoletedBy / Obsoletes            | EISSN                  |
| IsIdenticalTo                        | Handle                 |
| IsDescribedBy / Describes            | IGSN                   |
| HasMetadata / IsMetadataFor          | ISBN                   |
| IsDocumentedBy / Documents           | ISSN                   |
| IsReviewedBy / Reviews               | ISTC                   |
| IsPartOf / HasPart                   | LISSN                  |
| IsPublishedIn                        | LSID                   |
| IsCompiledBy / Compiles              | PMID                   |
| IsVariantFormOf / IsOriginalFormOf   | PURL                   |
| IsDerivedFrom / IsSourceOf           | UPC                    |
| IsRequiredBy / Requires              | URL                    |
| IsContinuedBy / Continues            | URN                    |
|                                      | w3id                   |

# DataCite connection metadata



# Metadata functions

## **Descriptive metadata supports discovery and citation**

- Titles, creators, subjects, description, dates...

## **Connection metadata promotes interoperability**

- Related identifiers
- Name identifiers – for creators and contributors
- Affiliation identifiers – for creators and contributors
- Funder identifiers – for funders

**Openly available metadata**

# Metadata is registered

- Registration methods:
  - [DataCite Fabrica - Form](#)
  - [DataCite Fabrica - File Upload](#)
  - [DataCite REST API](#) and platforms that use it
  - [DataCite MDS API](#) and platforms that use it
- Formats:
  - JSON
  - XML
  - [Other formats](#)

# Metadata is stored and indexed

- When you register a DOI, metadata is included in the **DataCite Metadata Store**
- All metadata registered with DataCite is made available under a CC0 license
- Metadata for findable DOIs is directly accessible through:
  - Interfaces:
    - [DataCite Commons](#)
  - APIs:
    - [DataCite REST API](#)
    - [DataCite GraphQL API](#)
    - [OAI-PMH service](#)
  - [Content Negotiation](#)



# Metadata is processed

- When a DOI is registered or updated, it is processed for DataCite Event Data
- Events are created for:
  - All related identifiers
  - Creators with ORCID iDs ("is authored by" event)
- Certain types of events are counted:
  - citations and references
    - citations: IsCitedBy, IsReferencedBy, IsSupplementTo
    - references: Cites, References, IsSupplementedBy
  - versions: HasVersion, IsVersionOf
  - parts: HasPart, IsPartOf

# **Doing more with DataCite metadata**



# Metadata completeness

- Include metadata that is *as complete as possible*
- Use the Recommended and Optional properties
  - Especially Subject, Description, Date, Contributor, GeoLocation, Rights
- Use properties / sub-properties for connection metadata:
  - RelatedIdentifier
  - nameIdentifier for Creator and Contributor
  - affiliationIdentifier for Creator and Contributor
  - funderIdentifier for FundingReference

and use other PIDs!

# Relation to FAIR principles

More complete metadata...

- makes research outputs easier to **find**
- clarifies conditions of **access**
- promotes **interoperability** between systems
- communicates the context to promote **re-use**

**What's next for the metadata  
schema?**

- Draft open for public comment through October 17:
  - Request for Comments: [DataCite Metadata Schema 4.5: Request for Comments](#)
  - Blog post: <https://doi.org/10.5438/q34f-c696>
  - Full schema draft: [https://datacite-metadata-schema.readthedocs.io/en/4.5\\_draft/](https://datacite-metadata-schema.readthedocs.io/en/4.5_draft/)
- Proposal includes:
  - support for DOIs for instruments
  - a new resource type for pre-registrations and registration reports
  - adding identifiers for publishers
  - new Distribution property
  - clarifications to RelatedItem property
  - updated PhysicalObject definitions
- After feedback period closes, aiming to release Version 4.5 in early 2023



CONNECTING RESEARCH,  
IDENTIFYING KNOWLEDGE



[info@datacite.org](mailto:info@datacite.org)



[pidforum.org](http://pidforum.org)



[datacite.org](http://datacite.org)  
[blog.datacite.org](http://blog.datacite.org)



[support.datacite.org](http://support.datacite.org)  
[support@datacite.org](mailto:support@datacite.org)



[@datacite](https://twitter.com/datacite)



[DataCite](https://www.youtube.com/DataCite)



[@datacite](https://www.linkedin.com/company/datacite)