DataCite Best Practice Guide

Christiane Bayer, IT-Gruppe Geisteswissenschaften (LMU)
(https://orcid.org/0000-0003-3074-4222, v2)
Andreas Frech, Universitätsbibliothek der LMU
(https://orcid.org/0000-0002-1458-1163, v2+v3)
Vanessa Gabriel, Universitätsbibliothek der LMU
(https://orcid.org/0000-0002-2058-5160, v2+v3)
Sonja Kümmet, Universitätsbibliothek der LMU
(https://orcid.org/0000-0002-8954-0200, v1)
Stephan Lücke, IT-Gruppe Geisteswissenschaften (LMU)
(https://orcid.org/0000-0002-5853-1918, v1+2)
Johannes Munke, Leibniz Supercomputing Centre
(https://orcid.org/0000-0002-5031-9170, v2)
Markus Putnings, Universitätsbibliothek der FAU
(https://orcid.org/0000-0002-6014-9048, v2)
Jürgen Rohrwild, Universitätsbibliothek der FAU
(https://orcid.org/0000-0002-1167-0339, v2+v3)
Julian Schulz, Max Weber Stiftung - Deutsche Geisteswissenschaftliche Institute im Ausland
(https://orcid.org/0000-0003-4374-2680, v1+2)
Martin Spenger, Universitätsbibliothek der LMU
(https://orcid.org/0000-0002-8841-5985, v1+v2+v3)
Tobias Weber, Leibniz Supercomputing Centre
(https://orcid.org/0000-0003-1815-7041, v1)

Version 3.0 (03.05.2024)

Contents

If you have any questions about this DataCite Best Practice Guide, please contact forschungsda
ten@ub.uni-muenchen.de or ub-fdm@fau.de

DataCite Best Practice Guide

The DataCite Metadata Schema [external link] has become a de facto standard for describing
research data. Despite all of its efforts to standardize metadata allocation, this schema offers
a great deal of leeway and alternatives in detail. For example, it is optional to clearly identify
languages used in the metadata either via ISO 639-1 or ISO 639-2 language tags. The value of
metadata increases with its consistency, which is achieved through compliance with standards.
One of the aims of the Best Practice Guide presented here is therefore to limit the choices provided
by DataCite by specifying a preference, and in this way to ensure as much consistency as possible.

This document is a guideline for the use of the official DataCite Metadata Schema documentation
[external link], version 4.5 [external link]. A support documentation for more convenience and
better navigation can be found here as a HTML version DataCite Metadata Schema Documentation
[external link, different versions available]. It is meant for researchers, IT and library support
staff. Further information on the schema can be found on the DataCite support site [external
link].

To create a DataCite XML file for the project you want to describe, we recommend to you to use the DataCite Metadata Generator [external link]. This tool is kept in sync with this guideline, safe for transmission times inbetween versions. If you want to create metadata for research data on a scale that is too large for manual procedures, please contact one of the institutions named above.

Overview

The first part, General Best Practice, is a selection of recommendations and obligations when using DataCite in general. It has been written in an FAQ-style (Frequently Asked Questions).

The second part, Best Practice for specific fields, gives more details for each of the 20 metadata fields of the DataCite metadata standard.

The third part, Examples, is a compilation of DataCite examples.

A. General Best Practice

- What do the metadata describe?
- What is the language of the metadata?
- How should I specify an institution?
- How should I specify a person?
- How should I handle different versions of research data?

B. Best Practice for specific fields

- 1 identifier [m]
- 2 creator [m]
- 3 title [m]
- 4 publisher [m]
- 5 publicationYear [m]
- 6 subject [m]*
- 7 contributor [r]
- 8 date [r]
- 9 language [o]
- 10 resourceType [m]
- 11 alternateIdentifier [o]
- 12 relatedIdentifier [r]
- 13 size [r]*
- 14 format [o]
- 15 version [o]
- 16 rights [m]*
- 17 description [m]*
- 18 geoLocation [r]
- 19 fundingReference [o]
- 20 relatedItem [o]

Mandatory fields are indicated by the tag [**m**], recommended fields by [**r**] and optional fields by [**o**]. Note: This guide deviates from the DataCite Metadata Schema 4.5 [external link] in the assessment of recommend and optional properties and assigns different levels of obligation to some of them. These are asterisked in the list above.

The presence of these fields improves discovery, makes long-term management of the datasets easier for the hosting institution and is helpful for future (re-)users of the dataset. The benefits outweigh the cost of requiring additional information. Furthermore, the information can often be provided by researchers in a straightforward way; like providing a short abstract in the *description*.

C. Examples

A. General Best Practice

What do the metadata describe?

Unless otherwise specified all information in the metadata concerns the research data (also denoted as "resource"), neither the project in whose context the data have been created or collected nor the metadata themselves.

What is the language of the metadata?

- The default language of the metadata is English. If another language is used, the same information must additionally be specified in English.
- Where language variations are possible (e.g. title, description, affiliations), the language should be specified by *xml:lang* attributes:

```
<title xml:lang="de">
        Bayerisches Musiker-Lexikon Online (BMLO)
</title>
<title xml:lang="en" titleType="TranslatedTitle">
        Digital Encyclopedia of Bavarian Musicians
</title>
```

- Proper nouns need not be translated.
- Use standardized data (e.g. controlled vocabularies) whenever possible. This might allow data aggregators to display the information in the language most suitable to the use case at hand.
- Recommendation: use either the two-letter language codes from ISO 639-1 or the three-letter language codes from ISO 639-2 (listed on Wikipedia [external link]). Be advised: the three-letter codes are used in library systems. If you use a different standard (i.e. BCP 47 [external link]), pay attention to be consistent and do not alternate between standards.

How should I specify a person?

- A person should be identified by name, persistent ID and affiliation
- It is recommended to not use titles/academic degrees in names as they are subject to change.
- State the name in the order "family name, given name". For example:

  ```
  <creatorName nameType="Personal">Krefeld, Thomas</creatorName>
  ```

4

- Recommendation: Additionally, separate family name and given name, each in a specific subfield:

```
<givenName>Thomas</givenName>
<familyName>Krefeld</familyName>
```

- Add a persistent identifier for persons, preferably a GND-ID (Gemeinsame Normdatei) [external link] or an ORCID-ID [external link] (Open Researcher and Contributor ID). This will make attributions robust to changes of names or affiliations:
    - Recommendation for GND-ID entries
        * GND-ID entries can be conveniently searched for on WebGND [external link] or lobid-gnd [external link]; for further search options see the GND website [external link]
        * Only use individualized GND entries that clearly identify a person (usually by year of birth and/or profession)

```
<nameIdentifier
    schemeURI="https://d-nb.info/gnd/"
    nameIdentifierScheme="GND">
        123778689
</nameIdentifier>
<nameIdentifier
    schemeURI="http://orcid.org/"
    nameIdentifierScheme="ORCID">
        0000-0001-9657-6052
</nameIdentifier>
```

- It is also recommended to indicate the affiliation to an institution (Note: An affiliation is an institution, not a project)
    - See How should I specify an institution
    - If a person has multiple affiliations:
        * It is recommended to state only one institution (the context of the resource determines the affiliation).
        * If unavoidable, multiple affiliations can be specified in the order of importance for the dataset published.

How should I specify an institution?

- Institutions may mainly be enterend in *affiliaton* or *fundingReference*
- Follow the policy of the institution.
- State the name of the institution as specific as possible (e.g. start with the chair/group, not with the university). If the name of the institution changed, use the name employed during the time the resource was created.
- Start with the more specific organizational units first and end with the most generic unit, separated by semicolon:

```
<affiliation xml:lang="de">
        Institut für Romanische Philologie; Ludwig-Maximilians-Universität München
</affiliation>

<affiliation
    xml:lang="de"
    affiliationIdentifier="https://ror.org/05591te55"
```

```
    affiliationIdentifierScheme="ROR">
        Ludwig-Maximilians-Universität München
</affiliation>
```

- If there is no policy or multiple names in multiple languages are given, use the English name.
- Always specify the language in which the name is given using a *xml:lang* tag.

```
<publisher xml:lang="en">Leibniz Supercomputing Centre</publisher>
<publisher xml:lang="de">Leibniz-Rechenzentrum</publisher>
```

- Affiliations are to be specified as of the time of creation of the resource.
- Add a persistent identifier (PID) for the institution, preferably a (ROR-ID [external link] (Research Organization Registry); if there is no entry in ROR, use a ISNI-ID [external link] (International Standard Name Identifier) or GND-ID.
- In the case of research funding organizations, it also makes sense to specify the CrossRef Funder Registry ID [external link].

```
<fundingReference>
    <funderName>Deutsche Forschungsgemeinschaft (DFG)</funderName>
    <funderIdentifier
        funderIdentifierType="Crossref Funder ID">
            http://dx.doi.org/10.13039/501100001659
    </funderIdentifier>
</fundingReference>
```

How should I handle different versions of the same research data?

Metadata can be updated without releasing a new version of the research data, but not vice versa; if the research data change, you need to update the metadata to reflect these changes.

If you want to publish several versions of the research data, but also want to have a point of reference for all of these publications together, we recommend to use a form of DOI-versioning [external link]:

- Specify a set of metadata that is valid for all versions.
- Specify a set of metadata for each version.
- Update all these metadata with the according references (e.g. include *isNewVersionOf* in the metadata of the new version, see *relatedIdentifier* for details).

B. Best practice for specific fields

1 identifier [m]

DataCite documentation [external link]

- This field can be omitted on submission: it is mandatory according to the DataCite standard, but it will be set by the data publisher.
- The assigned Digital Object Identifier (DOI) will be provided to you by the data publisher.

Example

```
<identifier identifierType="DOI">10.5282/ubm/data.158</identifier>
```

2 creator [m]

[DataCite documentation [external link]](#)

- This field is mandatory.
- Consult sections on [how to specify a person](#) and [how to specify a institution](#)
- Always prefer natural persons over institutions. If no natural person can be identified, you must set the attribute *nameType* to "Organizational".

Example

```xml
<creators>
    <creator>
        <creatorName nameType="Personal">Krefeld, Thomas</creatorName>
        <givenName>Thomas</givenName>
        <familyName>Krefeld</familyName>
         <nameIdentifier
            schemeURI="http://orcid.org/"
            nameIdentifierScheme="ORCID">
                0000-0001-9657-6052
        </nameIdentifier>
        <nameIdentifier
            schemeURI="https://d-nb.info/gnd/"
            nameIdentifierScheme="GND">
                123778689
        </nameIdentifier>
        <affiliation
            xml:lang="de">
                Institut für Romanische Philologie, Ludwig-Maximilians-Universität München
        </affiliation>
    </creator>
</creators>
```

3 title [m]

[DataCite documentation [external link]](#)

- This field is mandatory.
- Be as specific as you would be in the context of a journal publication.
- It is recommended to avoid filenames (e.g. "survey.csv") or generic descriptions (e.g. "survey data") as a title.
- The main title is specified without a *titleType*.
- The language of every title must be specified (see section on [metadata language](#) for the use of *xml:lang* attribute).
- If the main title is not specified in English, a title of type "TranslatedTitle" must be given in English.
- Title types "AlternativeTitle" and "Subtitle" are supported but not recommended. "Other" must not be used as a title type.

Example

```xml
<titles>
    <title
```

```
        xml:lang="de">
            Bayerisches Musiker-Lexikon Online (BMLO)
    </title>
    <title
        xml:lang="en"
        titleType="TranslatedTitle">
            Digital Encyclopedia of Bavarian Musicians
    </title>
</titles>
```

4 publisher [m]

DataCite documentation [external link]

- This field is mandatory.
- The language of the publisher name must be specified (see section on metadata language for the use of *xml:lang* attribute).
- If possible, the attibutes *publisherIdentifier*, *publisherIdentifierScheme* and *schemeURI* should be used.
- This field can be omitted on submission: it is mandatory according to the DataCite standard, but it will be set by the data publisher, i.e. the institution that hosts the (meta)data.

Example

```
<publisher
    xml:lang="de" publisherIdentifier="https://ror.org/05591te55"
    publisherIdentifierScheme="ROR" schemeURI="https://ror.org/">
        Universitätsbibliothek der Ludwig-Maximilians-Universität München
</publisher>
```

5 publicationYear [m]

DataCite documentation [external link]

- This field is mandatory.
- This field can be omitted on submission: it is mandatory according to the DataCite standard, but it will be set by the data publisher.

Example

```
<publicationYear>2019</publicationYear>
```

6 subject [m]

DataCite documentation [external link]

- This field is mandatory, in the DataCite standard it is only recommended.

Mandatory subject annotations

- The following subject annotations are mandatory (must occur at least once):

Type of Subject	Standard	Type of standard	Usage hint
Discipline	DDC	Classification	Use the English term for the discipline and include the three digit DDC notation via the *classificationCode* attribute (Canonical Source [external link])
Keywords	Wikidata QID and GND	Keyword	Wikidata and GND terms are both mandatory, including redundancy (if an appropriate entry does not exist contact the responsible Institution. Use Wikidata-Search [external link] and GND-Search [external link] to find the appropriate identifiers.

- It is also mandatory to include at least the *valueURI* or the *classificationCode* attribute.
- To improve machine-readablility we recommend using **both** *valueURI* and *classificationCode*.

Example

```
<subjects>
<!-- discipline specification using Dewey Decimal Classification (DDC) -->
    <subject
        xml:lang="en"
        subjectScheme="DDC"
        classificationCode="521">
        Celestial mechanics
    </subject>
<!-- keywords -->
    <subject
        xml:lang="en"
        subjectScheme="Wikidata QID"
        schemeURI="https://www.wikidata.org/wiki/"
        valueURI="https://www.wikidata.org/wiki/Q223776"
        classificationCode="Q223776">
        gravity assist
    </subject>
    <subject
        xml:lang="en"
        subjectScheme="GND"
        schemeURI="https://d-nb.info/gnd/"
        valueURI="https://d-nb.info/gnd/1135686874"
        classificationCode="1135686874">
        Gravity Assist
    </subject>
</subjects>
```

There should be no overlap between the discipline specifier(s) and the keywords.

Geotagging Specifying the location via subject is mandatory, if applicable to the resource:

- Canonical source for geonames is the GeoNames Service [external link] (a registration for API access is necessary).
- See *geoLocation* section for a more detailed specification.

Additional subject annotations

- Every other subject may be added.
- Specify the language of the subject.
- It is recommended to always qualify subjects by URL or scheme name. A good starting point to research existing schemes is BARTOC.org [external link] - Basic Register of Thesauri, Ontologies & Classifications. Unqualified subjects (not controlled by a controlled vocabulary, ontology or any other standard for the subject terms) are often useless for research data aggregators due to ambiguities.

Example

For this example a complete DataCite metadata file is available, see VerbaAlpina.

```xml
<subjects>
<!-- mandatory-->
    <subject xml:lang="en"
             subjectScheme="DDC"
             classificationCode="410">
                Linguistics
    </subject>
    <subject xml:lang="en"
             subjectScheme="DDC"
             classificationCode="004">
                Data processing computer science
    </subject>
    <subject
        xml:lang="de"
        subjectScheme="GND"
        schemeURI="https://d-nb.info/gnd/"
        valueURI="https://d-nb.info/gnd/4740815-7"
        classificationCode ="4740815-7">
          Chalet
    </subject>
    <subject
        xml:lang="en"
        subjectScheme="wikidata"
        schemeURI="https://www.wikidata.org/wiki/"
        valueURI="https://www.wikidata.org/wiki/Q136689"
        classificationCode="Q136689">
          chalet
    </subject>
    <subject
        xml:lang="fr"
        subjectScheme="wikidata"
        schemeURI="https://www.wikidata.org/wiki/"
        valueURI="https://www.wikidata.org/wiki/Lexeme:L643765"
```

```xml
            classificationCode="L643765">
                chalet
        </subject>
        <!-- optional-->
        <subject
            xml:lang="en"
            subjectScheme="Glottocode"
            schemeURI="https://glottolog.org/resource/languoid/id/"
            valueURI="https://glottolog.org/resource/languoid/id/high1286"
            classificationCode="high1286">
              High German
        </subject>
        <subject
            xml:lang="de"
            subjectScheme="geonames"
            schemeURI="http://www.geonames.org/"
            valueURI="http://www.geonames.org/2764958"
            classificationCode="2764958">
              Hall in Tirol
        </subject>
    </subjects>
</subjects>
```

7 contributor [r]

[DataCite documentation [external link]](#)

- This field is recommended if the data are published with a free license.
- If the license specified via the *rights* field restricts the usage in a way that possibly necessitates interaction with the rights holder, a *contributor* of type "RightsHolder" must be specified. Examples of free licenses are CC-0, CC-BY, or CC-SA; non-free licenses are for example CC-NC or CC-ND.
- Consult the sections on how to specify a person and how to specify a institution.
- If contributors change over versions, the version metadata should only include the actual contributors of the updated version. A metadata set representing all versions of the dataset (including links to the versions) can include all contributors with the dates of participation, see how to handle different versions of the research data.
- Duplicate mentions between *creator* and *contributor* are unproblematic.
- If a person has multiple roles, it is recommended to identify the most important role of that person and select only one.
- Be as specific as possible (a "ProjectLeader" is also considered to be a "ProjectMember", but "ProjectLeader" carries more information). Use generic role descriptions only when nothing else fits.
- The following roles are recommended:

Option	Description from DataCite standard (italics) and usage hints
ContactPerson	Person with knowledge of how to access, troubleshoot, or otherwise field issues related to the resource.
DataCollector	Person/institution responsible for finding, gathering/collecting data under the guidelines of the author(s) or Principal Investigator (PI).

Option	Description from DataCite standard (italics) and usage hints
DataCurator	Person tasked with reviewing, enhancing, cleaning, or standardizing metadata and the associated data submitted for storage, use, and maintenance of data quality.
DataManager	Person or organization responsible for digital maintainance of the finished resource, e.g. migration to new hardware, software and security updates for servers, access rights management.
Distributor	Institution responsible for dissemination of electronic or printed copies of the resource. The distributor is not neccessarily also a hosting institution of a digital resource, e.g., if server hosting is outsourced but the distributor still organizes access to the resource.
Editor	A person who oversees the details related to the publication format of the resource.
HostingInstitution	Typically, the organization allowing the resource to be available on the internet through the provision of its hardware/software/operating support.
ProjectLeader	Person officially designated as head of project team or sub-project team instrumental in the work necessary to development of the resource.
ProjectManager	Person officially designated as manager of a project. Project may consist of one or many project teams and sub-teams.
ProjectMember	Person on the membership list of a designated project/project team. All persons with a contract in the context of the project which produced the resource.
Researcher	A person involved in analyzing data or the results of an experiment or formal study. May indicate an intern or assistant to one of the authors who helped with research but who was not so "key" as to be listed as an author / creator.
ResearchGroup	Typically refers ot a group of individuals within a lab, department or division that has a specifically defined focus of activity.
RightsHolder	Person or institution owning or managing property rights, including intellectual property rights over the resource. Mandatory for non-free licenses; person or institution that owns the rights listed in field *rights*.
Sponsor	Organization or person that issued a contract or under the auspices of which a work has been printed, published, developed, etc.
Supervisor	We recommmed using this role for PhD advisors of the creators, who did not particiate as creators or in other roles themselves.
WorkPackageLeader	Person responsible for ensuring the comprehensive contents, versioning, and availability of the work package during the development of the resource.

Example

```xml
<contributors>
    <contributor contributorType="ProjectLeader">
        <contributorName nameType="Personal">Ludwig, Ralf</contributorName>
        <givenName>Ralf</givenName>
        <familyName>Ludwig</familyName>
        <nameIdentifier
            nameIdentifierScheme="ORCID"
            schemeURI="http://orcid.org/">
                0000-0002-4225-4098
```

```
            </nameIdentifier>
            <affiliation xml:lang="de">
                Department für Geographie, Ludwig-Maximilians-Universität München
            </affiliation>
        </contributor>
    <contributor contributorType="RightsHolder">
            <contributorName nameType="Personal">Štědronská, Markéta</contributorName>
            <givenName>Markéta</givenName>
            <familyName>Štědronská</familyName>
            <nameIdentifier
                nameIdentifierScheme="GND"
                schemeURI="https://d-nb.info/gnd/">
                    141321350
            </nameIdentifier>
            <affiliation xml:lang="de">
                Institut für Musikwissenschaft, Universität Wien
            </affiliation>
        </contributor>
</contributors>
```

8 date [r]

[DataCite documentation [external link]](#)

- This field is recommended.
- It is recommended to indicate date and time following the [W3C time and data formats [external link]](#). If the time is specified always include the time zone.

- Time periods can be specified by specifying the start date and the end date separated by a slash (/).
- The following types should be filled-out by the data producer:
    - Collected: time range when the resource was arranged (not necessarily identical to the time range when the resource was created).
    - Created: first version of a resource, cannot be identical with updated.
    - Updated: if the resource is a more recent version, cannot be identical with created.
- The following types are set by the publisher:
    - Submitted: point in time when the data were recieved by the data publisher.
    - Accepted: point in time when the data publisher decides to publish the data.
    - Issued: long format of the field *publicationYear*, point in time when a publisher publishes the data; should be set.
    - Available: only use in the context of embargo periods (this is not recommended).
    - Withdrawn: point in time when the publisher retracts the data publication.
- For dates describing the period the resource covers use "Other" for *dateType* and add "coverage" as a description under *datesInformation*, see example below.
- It is recommended to use the free text attribute *dateInformation* for disambiguation, if multiple dates with the same type are specified.
- "Copyrighted" as a *dateType* should not be used.

Example

```
<dates>
```

```xml
    <date dateType="Created">2016</date>
    <date
        dateType="Other"
        dateInformation="coverage">
            2050-09-01T00:00:00+01:00/2050-09-30T23:59:59+01:00
    </date>
</dates>
```

9 language [o]

DataCite documentation [external link]

- This field is optional.
- The field describes the main language of the **resource**, not of the metadata.
- Recommendation: use either the two-letter language codes from ISO 639-1 or the three-letter language codes from ISO 639-2 (listed on Wikipedia [external link]). Be advised: the three-letter codes are used in library systems.

Example

```xml
<language>en</language>
```

10 resourceType [m]

DataCite documentation [external link]

- This field is mandatory.
- DateCite allows various resource types.
- There are three goups of resources described by the metadata: Objects and instruments, discursive text, and research data.
- Decision tree to pick the right *resourceTypeGeneral*:

1. If you describe a physical object (biological sample, fragment of a meteorite) or an instrument (a book scanner, a microscope) use "PhysicalObject" and "Instrument", respectively. If not, proceed with 2.

2. Decide if the resource is data or discursive text (e.g. journal article or analytical text). If it's discursive text, choose one of the following:

   - Book
   - BookChapter
   - ConferencePaper
   - ConferenceProceeding
   - Dissertation
   - Journal
   - JournalArticle
   - OutputManagementPlan [Note: A data management plan is a special form of a output management plan]
   - PeerReview
   - Preprint
   - Report
   - Standard
   - StudyRegistration if not: Proceed with 3.

3. If the data submission contains heterogeneous data, consider publishing it in separate data publications or (less preferred) use "Collection". If the data are homogeneous, proceed with 4.

4. If the data are movies, images or sound files use "Audiovisual", "Image" or "Sound", respectively. If not, proceed with 5.

5. If the data are a digital, interactive representations of some real-world phenomena (e.g. trained models in the context of machine learning) use "Model". If not, proceed with 6.

6. If the data are descriptions of a workflow (e.g. in the common workflow language), use "Workflow". If not, proceed with 7.

7. If the data are an interactive resource like a virtual notebook use "ComputationalNotebook". If not, proceed with 8.

8. If the data are source code files (incl. configuration and built artefacts), use "Software". If not, proceed with 9.

9. If the data have a fixed structure (e.g. table-like), use "Dataset". If not, proceed with 10.

10. If the data are text files, use "Text". If not, proceed with 11.

11. Use "Other".

- These *ResourceTypes* should **not** be used:
  - DataPaper
  - Event
  - InteractiveResource
  - Service
- Note: Only items with the *resourceTypeGeneral* "Dataset" will be included in the Google Dataset Search. All other types are currently not supported.

Examples

```
<resourceType resourceTypeGeneral="Dataset">
    Regional Climate Measurements
</resourceType>

<resourceType resourceTypeGeneral="OutputManagementPlan">
    Data Management Plan
</resourceType>
```

11 alternateIdentifier [o]

DataCite documentation [external link]

- This field is optional.
- These alternate IDs additionally identify the resource, meaning that it can also be found via these identifiers and distinguished from other resources by this ID.
- The *alternateIdentifier* can be a persistent, globally unique ID. However, the field may also be used for identifiers, which are only unique and specific in the context of the research project (e.g. local identifiers or workspace identifiers) but not globally. Examples for alternate identifiers are sequence numbers, time stamps or database numbers. Contrary, the global identifier in field *identifier* **must** be a DOI.
- The attribute *alternateIdentifierType* must be used to specifiy the type of the identifer.

Recommendation for *alternateIdentifierType*:

For common global identifers, just specify the name of the identifier or its acronym. Examples of such identifiers are: ARK, arXiv, bibcode, DOI, EAN13, EISSN, ePIC, Handle, IGSN, ISBN, ISSN, ISTC, LISSN, LSID, PMID, PURL, UPC, URL, URN, and w3id.

For other identifiers we recommend to first give the origin of the ID:

- project-specific identifier: an ID that has meaning inside the project that created the data.
- application-specific identifier: an ID that has meaning in the context of an application that is used to process the data.
- institution-specific identifier: an ID that has meaning in the context of the institution that provides, funded or created the data.

and then add, separated by a slash (/), the name of the identifer, if known. This way, even if the name of the ID is relatively obscure, the broader context of the ID can still be identified.

Example

Each VerbaAlpina dataset is assigned an internal ID [external link] (a project-specific identifier) as well as a persistent LMU-UB ID [external link] (in short lmUB - an institution-specific identifier) by the data repository.

```
<alternateIdentifiers>
    <alternateIdentifier alternateIdentifierType="institution-specific identifier/lmUB">
        68fd5294-9077-3983-a20e-7f25c074c4c7
    </alternateIdentifier>
    <alternateIdentifier alternateIdentifierType="project-specific identifier/VA-ID">
        L91_v8
    </alternateIdentifier>
    </alternateIdentifiers>
<relatedIdentifiers>
```

12 relatedIdentifier [r]
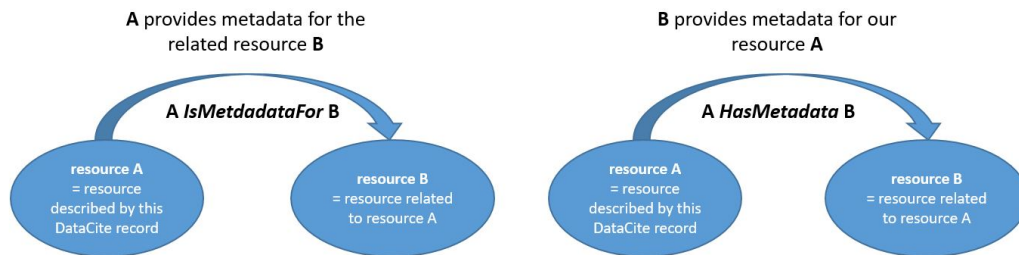
DataCite documentation [external link]

Note that DataCite provides two separate elements to establish relationships between resources: *relatedIdentifier* and *relatedItem*.

The *relatedIdentifier* element should be used for machine-readable identifiers (like DOI, PubMed ID or ISBN). This identifier points to additional information on the related resource. It is recommended as it facilitates automated discovery of the described resource.

The element *relatedItem* can be used for information on a related object that does not have an identifier (e.g. conference contributions or book series). It is also useful if an identifier exists but one wants to provide additional, more specific information on the related resource, e.g. to provide the page number, volume and title of a journal.

- This field is recommended.
- If possible, relations of the described resource should be mirrored in the metadata of the related resource. For example, make sure that a paper referencing a dataset includes the identifier of the dataset in its metadata and vice versa. This also applies to all reciprocal *relationTypes* (e.g. "IsNewVersionOf" and "IsPreviousVersionOf").

- The *relationIdentifierType* must be taken from a fixed list [external link]: ARK, arXiv, bibcode, DOI, EAN13, EISSN, Handle, IGSN, ISBN, ISSN, ISTC, LISSN, LSID, PMID, PURL, UPC, URL, URN, w3id.
- The publisher may curate the list of *relatedIdentifiers* (e.g. adding relevant related resources over time on a best effort basis).
- The relations are always specified from the perspective of the described resource (refered to as resource A in the examples). The related resource is called resource B in the examples. For detailed definitions of all relationTypes see the DataCite schema documentation [external link]

A provides metadata for the related resource **B**

A *IsMetdadataFor* B

resource A
= resource
described by this
DataCite record

resource B
= resource related
to resource A

B provides metadata for our resource **A**

A *HasMetadata* B

resource A
= resource
described by this
DataCite record

resource B
= resource related
to resource A

- Only use these *relationTypes*:

relationTypes

Relation categories	Usage	relationType	Guidance / Example
Versions	Use to differentiate specific and unspecific versions of a resource	HasVersion (IsVersionOf); IsNewVersionOf (IsPreviousVersionOf); IsVariantFormOf (IsOriginalFormOf); IsIdenticalTo; Obsoletes (IsObsoletedBy)	Be as specific as possible. Use "NewVersion" and "PreviousVersion" if A is the predecessor or successor of B; use of "Obsolets" and "IsObsoletedBy" is recommended for standards, legal regulations, etc. If the version is unspecific use "HasVersion" and "IsVersionOf".
Hierarchy	Use to create hierarchical relationships	IsPartOf (HasPart)	"HasPart" indicates A includes the part B; "IsPart" indicates A is a portion of B Example: a container "HasPart" a dataset

Relation categories	Usage	relationType	Guidance / Example
Provenance	Use to refer to source materials (for software also see below)	IsSourceOf; IsDerivedFrom; Continues (IsContinuedBy); Collects (IsCollectedBy)	Examples:Volume 2 of a book series (resource A) "Continues" volume 1 (resource B) A dataset (resource A) "IsSourceOf" a diagram (resource B) An Instrument (resource A), e.g. a microscope, "Collects" an image set (resource B); a PhysicalObject (resource A) "isCollectedBy" an Instrument (resource B)
Complementary (for software see below)	Use for resources that build on or complement each other	IsSourceOf; IsDerivedFrom; Continues (IsContinuedBy); IsSupplementTo (IsSupplementedBy)	Examples:Volume 2 of a book series (resource A) "Continues" volume 1 (resource B). A computer notebook (resource A) "IsSupplementTo" an article (resource B)
Bibliographic	Use to relate published texts/material	IsPublishedIn; References (IsReferencedBy); Cites (IsCitedBy)	Examples:An article (resource A) "IsPublishedIn" an edited volume (resource B), e.g., conference proceedings. A presentation (resource A) "References" a dataset (resource B) when discussing a plot.
Additional information	Use for resources that provide further information on the described resource	Documents (IsDocumentedBy); HasMetadata (IsMetadataFor); Describes (IsDescribedBy); Reviews (IsReviewedBy)	A digtital representation of a painting (resource A) "HasMetadata" in Europeana (Resource B) An article (resource A) "isDescribedBy" a PubMed entry (resource B)
Software specific	IsRequired / Requires can be used to indicate software dependencies. IsCompiledBy and Compiles relate software code and compiler	Requires (IsRequiredBy); Compiles (IsCompiledBy)	A piece of code (resource A) "Requires" a software library (resource B)

Example 1 (DOI)

The ClimEx Project "IsDescribedBy" an article in the Journal of Applied Meteorology and Climatology. The article's DOI is 10.1175/JAMC-D-18-0021.1 . For the full DataCite metadata for this example see ClimEx below.

```xml
<relatedIdentifiers>
    <relatedIdentifier relatedIdentifierType="DOI" relationType="IsDescribedBy">
        10.1175/JAMC-D-18-0021.1
    </relatedIdentifier>
</relatedIdentifiers>
```

Example 2 (PubMed)

PubMed provides a description and additional information on an article published in the journal Nature. The article "IsDescribedBy" a PubMed entry.

```xml
<relatedIdentifiers>
    <relatedIdentifier
        relatedIdentifierType="PMID"
        relationType="IsDescribedBy">
            34552256
    </relatedIdentifier>
    <relatedIdentifier
        relatedIdentifierType="arXiv"
        relationType="IsNewVersionOf">
            arXiv:2107.02222
    </relatedIdentifier>
</relatedIdentifiers>
```

Example 3 (Wikidata)

Wikidata provides metadata for the BMLO project and dataset. Thus, BMLO "HasMetadata" in Wikidata under Q-ID 47191.

```xml
<relatedIdentifiers>
    <relatedIdentifier
        relatedIdentifierType="URL"
        relationType="HasMetadata"
        subjectScheme="wikidata"
        schemeURI="https://www.wikidata.org/wiki/">
            Q47191
    </relatedIdentifier>
</relatedIdentifiers>
```

Example 4 (Pangaea)

Pangaea publishes datasets in the field of Earth & Environmental Science. Pangaea shows the sources that were used during the creation of the research dataset, similar to a list of references. For the full DataCite metadata for this example see DataCite Search: https://commons.datacite .org/doi.org/10.1594/pangaea.941445 .

```xml
<relatedIdentifiers>
    <relatedIdentifier relationType="References"
```

```
        relatedIdentifierType="DOI">10.1016/0034-6667(75)90049-4</relatedIdentifier>
    <relatedIdentifier relationType="References"
        relatedIdentifierType="DOI">10.1016/j.revpalbo.2020.104236</relatedIdentifier>
    <relatedIdentifier relationType="References"
        relatedIdentifierType="DOI">10.1016/j.revpalbo.2019.02.004</relatedIdentifier>
    <relatedIdentifier relationType="References"
        relatedIdentifierType="DOI">10.1080/01916122.2014.940472</relatedIdentifier>
    <relatedIdentifier relationType="References"
        relatedIdentifierType="DOI">10.1191/095968398671104653</relatedIdentifier>
    <relatedIdentifier relationType="References"
        relatedIdentifierType="Handle">1885/144170</relatedIdentifier>
</relatedIdentifiers>
```

13 size [r]

[DataCite documentation [external link]](#)

- This field is recommended, whereas it is optional in the DataCite standard.
- This field is repeatable. Thus, different measures for the size / volume of the dataset can be given.
- If you make use of this field, always specify the size in Bytes (denoted by 'B' - note that a lower case 'b' stands for bit). Prefered are: kB, MB, GB, TB etc. Separate number and unit with one space. The decimal separator must be the decimal point, e.g., *7.23 GB*.
- If the data are compressed, specify the size of the compressed file/archive.
- If the data consist of several units (without using an archival software), specify their combined sizes.
- Further information on the data size (e.g. runtime of an audio file or number of images) can be given in a separate *size* field as free text. Note that such information can also be given in the *description*.

Example

```
<sizes>
    <size>7.23 GB</size>
    <size>34 min (length audio file)</size>
</sizes>
```

14 format [o]

[DataCite documentation [external link]](#)

- This field is optional.
- Use MIME type format as specified in [RFC 2646 [external link]](#), possible values should be taken from [the IANA list of Media Types [external link]](#).
- Specify in this order (skip if it does not apply):

1. If files are compressed, append the MIME type of the compressed file to the MIME type of the uncompressed file using a "+" sign (e.g. text/xml+zip).
2. If files are in an archive, specify the MIME type of the archive format, for example "application/tar". This information is useful to determine in advance which software tools are needed to access the archived files.

3. Specify each MIME type in a separate field in alphabetical order, do not repeat MIME types.

Example

```xml
<formats>
    <format>application/tar+gzip</format>
    <format>application/netcdf</format>
    <format>text/plain</format>
    <format>text/csv</format>
</formats>
```

15 version [o]

DataCite documentation [external link]

- This field is optional.
- Note that this field refers to the **version of the resource**, not the version of the metadata.
- The versioning information is set according to the policies of the data provider (data publishers do not change/use this field).
- It is recommended to use semantic versioning [external link]. Up to three labels are supported (Major, Minor, Patch). Depending on the resource, only one or two labels might be needed.

Example 1 (two labels - Major.Minor)

```xml
<version>4.2</version>
```

Example 2 (three labels - Major.Minor.Patch)

```xml
<version>4.2.1</version>
```

16 rights [m]

DataCite documentation [external link]

- This field is mandatory, whereas the DataCite standard specifies it as optional.
- If applicable, *rightsURI* must be set.
- To avoid inconsistencies only assign a single license to the described dataset or the described software code.
- It is not recommended to publish both research data and software code as part of a single publication (consider two separate publications, see *resourceType*).

Guidance for using a license:

- Recommendation: Creative Commons (CC) [external link] as license for data and Apache 2.0 license [external link] for software.
- Use the standardized short identifier list provided by SPDX [external link] to specify the license in the *rightsIdentifier* attribute.
- You should not use CC licenses with the NC or ND limitation to ensure reusability (although submissions with these limitations are accepted).
- For further license options consult Choose a license [external link] or the CC license helper [external link].

Example 1 (Dataset)

21

```
<rightsList>
    <rights
        xml:lang="en-US"
        schemeURI="https://spdx.org/licenses/"
        rightsIdentifierScheme="SPDX"
        rightsIdentifier="CC0-1.0"
        rightsURI="http://creativecommons.org/publicdomain/zero/1.0/">
            CC0 1.0
    </rights>
</rightsList>
```

Example 2 (Software)

```
<rightsList>
    <rights
        xml:lang="en-US"
        schemeURI="https://spdx.org/licenses/"
        rightsIdentifierScheme="SPDX"
        rightsIdentifier="Apache-2.0"
        rightsURI="https://www.apache.org/licenses/LICENSE-2.0">
            Apache License 2.0
    </rights>
</rightsList>
```

17 description [m]

[DataCite documentation [external link]](#)

- This field is mandatory, whereas the DataCite standard only recommends it: There has to be at least one entry of type "Abstract" in English.
- Always specify the used language (*xml:lang* attribute) of each description.
- If there are descriptions in more than one language, the content may be different (no literal translation required).
- Each description has a limit of 300 words.
- Description of *descriptionType* "Methods" is optional. Best practice: use keywords from this controlled list, separated by comma.
- Description of *descriptionType* "TechnicalInfo" is optional. Best practice: use keywords from this controlled list, separated by comma. Additionally, data producers could consider creating a README file and link it via the *relatedIdentifier* field.
- These types are not recommended:
    - SeriesInformation (If needed, information on series title, volume, issue, or page number should be provided via the new *relatedItem* field.)
    - TableOfContents
    - Other

Example

```
<descriptions>
    <description xml:lang="en" descriptionType="Abstract">
        The "Kritische Ausgabe der Werke von Richard Strauss", a long-term editorial
        project, has been under way at the Institut für Musikwissenschaft of the
        Ludwig-Maximilians-Universität Munich since 2011; it is directed by ...
```

```xml
    </description>
    <description xml:lang="de" descriptionType="Abstract">
        Das Langzeit-Editionsprojekt „Kritische Ausgabe der Werke von Richard Strauss"
        wird seit Februar 2011 unter der Leitung von Prof. Dr. Hartmut Schick am
        Institut für Musikwissenschaft der Ludwig-Maximilians-Universität München ...
    </description>
    <description xml:lang="en" descriptionType="Methods">
        digital editing, software/application development
    </description>
</descriptions>
```

18 geoLocation [r]

[DataCite documentation [external link]](#)

- This field is recommended.
- Describes the resource (e.g. where an image has been taken or where a sensor is located),
  not the related project or institute, if the former is not applicable, do **not** use it for the
  latter.
- *geoLocationPlace* must be identical to corresponding GeoNames field in the *subjects*, consult
  the geotagging subsection.
- Canonical source for coordinates is the GeoNames Service [external link].

Examples

- *geoLocationPlace* and *geoLocationPolygon*:

```xml
<geoLocations>
    <geoLocation>
      <geoLocationPlace>Höslwang</geoLocationPlace>
      <geoLocationPolygon>
        <polygonPoint>
          <pointLatitude>47.9231796264648</pointLatitude>
          <pointLongitude>12.2860469818115</pointLongitude>
        </polygonPoint>
        <polygonPoint>
          <pointLatitude>47.9231796264648</pointLatitude>
          <pointLongitude>12.3512439727784</pointLongitude>
        </polygonPoint>
        <polygonPoint>
          <pointLatitude>47.9707412719727</pointLatitude>
          <pointLongitude>12.3512439727784</pointLongitude>
        </polygonPoint>
        <polygonPoint>
          <pointLatitude>47.9707412719727</pointLatitude>
          <pointLongitude>12.2860469818115</pointLongitude>
        </polygonPoint>
        <polygonPoint>
          <pointLatitude>47.9231796264648</pointLatitude>
          <pointLongitude>12.2860469818115</pointLongitude>
        </polygonPoint>
```

```xml
            </geoLocationPolygon>
        </geoLocation>
    </geoLocations>
```

- *geoLocationPlace* and *geoLocationBox*:

```xml
<geoLocations>
    <geoLocation>
        <geoLocationPlace>Hall in Tirol</geoLocationPlace>
        <geoLocationPoint>
            <pointLongitude>11.51667</pointLongitude>
            <pointLatitude>47.28333</pointLatitude>
        </geoLocationPoint>
        <geoLocationBox>
            <westBoundLongitude>11.5272636413574</westBoundLongitude>
            <eastBoundLongitude>11.4707803726196</eastBoundLongitude>
            <southBoundLatitude>47.2697830200196</southBoundLatitude>
            <northBoundLatitude>47.2893867492676</northBoundLatitude>
        </geoLocationBox>
    </geoLocation>
</geoLocations>
```

19 fundingReference [o]

DataCite documentation [external link]

- This field is optional.
- This is the place to add information about the project and its funding.
- *funderName* is mandatory, if *fundingReference* is used. For usage see How should I Specify an Institution.
- Use Cordis (EU) [external link], Gepris (DFG) [external link], FWF (Austria) [external link] et al. for identification of grants.
- *awardTitle* is the name of the grant, not the funding line or funding program.

Example

```xml
<fundingReferences>
    <fundingReference>
        <funderName>Deutsche Forschungsgemeinschaft (DFG)</funderName>
        <funderIdentifier funderIdentifierType="ROR">
            https://ror.org/018mejw64
        </funderIdentifier>
        <awardNumber awardURI="http://gepris.dfg.de/gepris/projekt/253900505">
          253900505</awardNumber>
        <awardTitle xml:lang="de">
            VerbaAlpina. Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit
        </awardTitle>
    </fundingReference>
</fundingReferences>
```

20 relatedItem [o]

Note that DataCite provides two separate elements to establish relationships between resources: *relatedItem* and *relatedIdentifier*.

relatedItem complements the field *relatedIdentifier*. Only use *relatedItem* if

1. the related resource you want to add does not have a PID/DOI to use *relatedIdentifier* or
2. you want to provide addional context for a published resource (e.g. page numbers or the number of the chapter in a book), as such information can only be provided with *relatedItem* and not with *relatedIdentifier*.

- This field is optional.
- Information given in this field is not evaluated automatically. The information stored in *relatedItem* is not used in the PID graph and EventData. If a *relatedItemIdentifier* is provided, an identical *relatedIdentifier* is strongly recommended for indexing.
- *relatedItem* uses the same *relationTypes* as *relatedIdentifier*.

Example

Here information on a journal series is provided, in particular the journal title, volume, issue, page numbers as well as the publisher.

```xml
<!-- Article is published in Journal "Ladinia"; Article = A, Journal = B -->
<relatedItems>
    <relatedItem relatedItemType="Journal" relationType="IsPublishedIn">
        <relatedItemIdentifier relatedItemIdentifierType="ISSN">
      1124-1004</relatedItemIdentifier>
      <titles>
        <title xml:lang="de">Ladinia</title>
        <title titleType="AlternativeTitle" xml:lang="de">
          Revista scientifica dl Istitut Ladin Micurá de Rü</title>
      </titles>
      <publicationYear>2019</publicationYear>
      <volume>43</volume>
      <issue>1</issue>
      <firstPage>139</firstPage>
      <lastPage>155</lastPage>
      <publisher>Istitut Ladin Micurá de Rü</publisher>
    </relatedItem>
</relatedItems>
```

Examples

Digital encyclopedia: "Bayerisches Musiker-Lexikon Online"

The Digital Encyclopedia of Bavarian Musicians (Bayerisches Musiker-Lexikon Online, short BMLO) refers as a musicological model project. The BMLO offers a digital biographical dictionary focussing on music science, furthermore it enriches the presented personalities of Bavarian history of music by implementing further information, gathered from biographical literature, archives, libraries and digital collections. In this way, the BMLO constitutes the core of an interconnected, virtual cluster for history of music. Currently, 24621 out of a total of 27818 records is presented

on the web. As parts of this semantic network should also be mentioned the Munich Dictionary of Musik (Münchner Musiklexikon, short MUK), which serves since 2010 as a encyclopedia for music corporations with a linkage to Munich, as well as LOCI, a geographic database for music, culture and history, founded in 2012.

BMLO metadata example [external link]

Meteorological project: "ClimEx"

The ClimEx project investigates the effects of climate change on meteorological and hydrological extreme events and implications for water management in Bavaria and Québec. It especially consists of two new aspects:

- An ensemble of 50 transient runs of the canadian general circulation model CanESM2 (~200km resolution) from 1950 to 2100, resulting in 7500 years of modelled climate. As each of these runs is initialized with only slightly altered starting conditions, this ensemble can be interpreted as (modelled) natural variability. The CanESM2 then drives the regional climate model CRCM5 (~11km resolution) for a domain that covers most of central Europe. Both models are internationally established and widely used in the climate science community.

- A physically based hydrological model (WaSiM) is driven by this climate input for the entire hydrological Bavaria at very high temporal and spatial resolution of 3 hours and 500m to investigate both climate change impacts and natural variability of extreme events, especially floods.

ClimEx further strengthens the international collaboration between Bavaria and Québec as research facilities, universities and public water agencies intensify their former cooperation approaches.

ClimEx metadata example [external link]

Volume as part of a series: "Discourses on Corruption"

The volume "Discourses on Corruption. Interdisciplinary and Intercultural Perspectives" is part of the series "Politics and Society in India and the Global South", a collaboration of Sage Publications and the M.S. Merian – R. Tagore International Centre of Advanced Studies 'Metamorphoses of the Political' (ICAS:MP). ICAS:MP is an Indo-German research collaboration of six Indian and German institutions funded by the German Federal Ministry of Education and Research (BMBF). Located in New Delhi, ICAS:MP critically intervenes in global debates in the social sciences and humanities. This volume, through case studies, investigates corruption in the Global South (especially India and Brazil) and West (especially Switzerland) to gain a more nuanced view of the phenomenon.

MWS metadata example [external link]

Article in a conference proceeding: "High-Energy Physics"

The article summarizes the contents of a talk given at the Workshop on Exclusive Reactions at High Momentum Transfer, 21-24 May 2007, in Newport News, VA, United States. The article deals with the radiative decay of delta baryons to nucleons and an on-shell photon. The form factors of the decay are dertermined using QCD light-cone sum rules and photon distribution amplitudes. The proceeding was published as an edited volume in printed and electronic form.

Furthermore, there are several works that are closely related to the talk: a journal publication and a preprint version of the article. These have been referenced in the metadata.

High-Energy Physics article [external link]

Critical editon (digital & print): "Richard Strauss Kritische Werkausgabe"

The "Kritische Ausgabe der Werke von Richard Strauss", a long-term editorial project, has been under way at the Institut für Musikwissenschaft of the Ludwig-Maximilians-Universität Munich since 2011; it is directed by Prof. Dr Hartmut Schick and is supervised by a project committee and advisory board constituted by the Bayerische Akademie der Wissenschaften. The project is part of the so-called Akademienprogramm, financed jointly by Germany's federal government and federal states. Collaborations are in place with the Richard-Strauss-Institut in Garmisch Partenkirchen – which, between 2009 and 2012, prepared the "Richard-Strauss-Quellenverzeichnis" (www.rsi-rsqv.de), funded by the Deutsche Forschungsgemeinschaft – with the IT-Gruppe Geisteswissenschaften at the LMU Munich, and the Richard-Strauss-Archiv at Garmisch Partenkirchen, run by the composer's family.

RSW metadata example [external link]

Digital lexicographical information system: "VerbaAlpina"

The project VerbaAlpina [external link], funded by the DFG from 2014 to 2023 as a long-term project [external link], was dedicated to the documentation of dialectal lexical variation in the Alpine region within concept domains typical of the region. An independent goal of the project was to use digital methods as extensively and consistently as possible. Problems resulting from complete digitality were discussed, reflected upon, documented and, where possible, solutions were found. Essentially, VerbaAlpina is a lexicographical information system that unites the two traditionally mutually exclusive publication genres of "dictionary" and "language atlas" in one. The results of the VerbaAlpina project are available exclusively in electronic form. Essentially, this is the highly granularly structured lexicographical data material, as well as explanatory texts of various orientations and software components developed as part of the project. The aforementioned electronic legacies of VerbaAlpina were treated in accordance with the FAIR criteria. In the course of this, data was also transferred to the research data repository [external link] of the LMU University Library.

VerbaAlpina metadata: example full data set [external link]; example individual data set [external link]