Version rc2 2022-07-20

# Minimal descriptive metadata for research data

## Introduction

### Definition of terms

Data and metadata standards can be complex and difficult to start working with. This document provides a practical guideline to working with a generic minimal metadata description that can be used to describe any dataset. In this document we define: - **dataset** as information generated, or used, that is conceptually and/or logically related and can be written down (encoded/serialised) as one or more files in a machine readable format. - **metadata** as information that describes a dataset and can be written down (serialised/encoded) in one or more standard formats and a - **standard format** a formally defined metadata schema that is well defined, serialisable and available in a machine readable format. In this document the metadata properties are derived from and map to a subset of the DataCite Metadata Schema.

### Why (minimal) metadata?

In principle, *richer metadata is better*, however, the properties used in this subset should be applicable to (and supported by) a wide a range of tools, platforms etc. In addition, the properties decscribed in this document should be usable by researchers who do not have a domain specific alternative and who, in general, may even be discouraged from using metadata at all if *required* to add extensive metadata to every dataset they produce. Instead a minimalist approach is used i.e. the number of *required* properties is kept to a mimnimum (only those necessary to ensure DataCite interoperability). Furthermore a more extensive set of properties is provided that allows for richer descriptions of the data as well as semantic linking to other resources. The use of these additional properties is highly recommended.

**For archiving** By archiving we mean storing an immutable copy of a dataset for a longer period (5 or 10 years). During such a period all the researchers involved in the particular project might have left the organization. This means there should be enough metadata to identify who created the dataset, what it is and why it was produced.

**For publishing** By publishing we mean sharing the metadata and, if possible, the dataset itself on the internet making it findable and reusable by other researchers. In this way, like DataCite, our minimal metadata guidelines enables FAIR (Findable, Accessible, Interoperable, Reusable) data management.

The VU data management policy requires published datasets to be registered in

the research information system Pure and rhe minimal metadata specification includes the necessary properties that allows for easy dataset registration in Pure. When research data is deposited in widely-used, registered, repositories (DataverseNL, Yoda, Zenodo, etc.) it can be automatically harvested and registered in Pure without having to enter the same information twice.

**Human vs. machine readability**

When describing a dataset think about how the metadata will be used. Computers will index the metadata so you should add relevant keywords, description and a title so your dataset will pop up in an internet search. Make sure you as creator and all the contributors are correctly named so published datasets are correctly attributed to you and automated systems can attribute the dataset to your research output. Explicitly adding persistent URLs to related publications or datasets to the metadata enables them to be efficiently linked together.

Once someone finds your dataset he or she will want to read your description to quickly see if the dataset is relevant, so it is important that the *Description* should be human readable. While the description should describe the data it is generally good practice to add extra information about the dataset in additional documentation. This can take the form of a README.txt file or codebook and can provide more context on how the data was gathered and processed, the experimental protocols and software used to generate the dataset as well as the filename system and variable nanes used in the individual files etc.

## How to use this document?

Most, if not all, repositories and publication platforms will use their own webform for metadata with mandatory and recommended fields. Consider this document as a guideline for recommended and mandatory fields extra to those the particular system requires. Sometimes properties might be differently named (for example authors vs creators) but in all cases it should be possible to enter, at least, the mandatory metadata that is advised in this document. Similarly, during your research, if domain-specific metadata exists, that includes the mandatory properties described here, then it is recomended to use the domain-specific format as metadata. However, in cases where no domain-specific metadata exists, the guidelines presented here should be considered.

If your storage system does not provide you with functionality for entering metadata (for example ResearchDrive, OneDrive) you can consider using the human-writable, machine readable form of these guidelines, Melite, which we have developed and can be saved as a plaing text file.

## Using related identifiers

This metadata specification allows you to link the dataset or collection that it describes to other online resources. More commonly know as Linked Data these

relations form the basis of the Semantic Web and can be thought of as a set of statements that relate a *subject* (the dataset described by the metadata) using a *predicate* (the related identifier property) to an *object* (represented by a unique identifier).

For example, consider metadata describing this specification document (the dataset), we could say: * the dataset is derived from DataCite 4. * the dataset is a version of the text that is developed on GitHub. * the dataset is the source of the human-readable format Melite

If we add in some specific detail and make the subject implicit (everything is about the dataset) we can rewrite the above as: * **IsDerivedFrom** 10.14454/3w3z-sa82 * **IsVersionOf** https://github.com/vu-rdm-tech/metadata * **IsSourceOf** https://github.com/vu-rdm-tech/melite-metadata

By doing this we have linked or mapped the relations between our document to other internet resources in a machine readable way.

## Properties and defined types

### Properties and their explanation

**M** Considered mandatory for findability of your dataset and correct registration in Pure
**R** Recommended for optimal findability
**O** Optional

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 1 | Identifier | | M | O | This should be a global unique identifier, which preferably is un-changeable and links to the datasets. In most cases the repository where you publish your data will generate this in the form of a Handle or DOI.If no persistent URL is available you could use a normal URL, but be aware that you should not move the data after-wards.(A persistent identifier is optional for unpub-lished archived datasets.) |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 2 | Creator(s) | | M | M | The main researchers involved in producing the data, in priority order. |
| 2a | | Name | M | M | Enter names of persons as: <family name>, <first name> <initials> e.g. Olivier, Brett G. |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 2b | | Affiliation(s) | M | M | Always make sure to always enter your affiliations when you archive or publish your dataset.Make sure you at least add the VU, the correct name for the VU is “Vrije Universiteit Amsterdam”.Note: some repositories may allow you to enter a ROR identifier (Research Organization Registry). The VU ROR is: https://ror.org/008xxew50 |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 2c | | Identifier(s) | R | R | If known, enter one or more unique identifiers like AuthorID, ORCID, ISNI or ResearcherID.The VU strongly recommends registering for an ORCID (https://orcid.org/). This is an easy way to uniquely identify yourself over which you have full control. |
| 3 | Title | | M | M | A descriptive title for your dataset, should not be longer than about 200 characters |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 4 | Publisher | | M | O | Name of the organization where you published your dataset. In most cases the repository where you upload your data will fill this in automatically. Otherwise fill in the name of the organization owning the website or database.(This field only applies to published datasets.) |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 5 | Publication Year | | M | O | The year (or date) you first published your dataset. the repository where you upload your data will usually generate this automatically.(This field only applies to published datasets.) |
| 6 | Subject(s) | | R | R | Provide a list of keywords describing your dataset. This will make it easier to find your dataset on the internet.Some repositories will have controlled term lists to choose from. |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 7 | Contributor(s) | | R | R | The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource.For software, if there is an alternate entity that "holds, archives, publishes, prints, distributes, releases, issues, orproduces" the code, use the contributor Type "hostingInstitution" for the code repository. |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 7a | | Name | M | M | Enter names of persons as: <family name>, <first name> <initials> e.g. Olivier, Brett G. |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 7b | | Affiliation(s) | M | M | Always make sure to always enter your affiliations when you archive or publish your dataset.Make sure you at least add the VU, the correct name for the VU is “Vrije Universiteit Amsterdam”.Note: some repositories may allow you to enter a ROR identifier (Research Organization Registry). The VU ROR is: https://ror.org/008xxew50 |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 7c | | Identifier(s) | R | R | If known, enter one or more unique identifiers like AuthorID, ORCID, ISNI or ResearcherID.The VU strongly recommends registering for an ORCID (https://orcid.org/). This is an easy way to uniquely identify yourself over which you have full control. |
| 7d | | Type | M | M | The role of the contributor, see the table for possible types. If contributor is used then Contributor Type is mandatory. |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 8 | Date(s) | | R | R | If applicable add extra dates applying to your dataset. A good addition is the “Date collected”, meaning the date or date range when you collected the dataset. |
| 9 | Language | | O | O | The primary language of your dataset.Please use a 2 letter code (e.g. en, nl, fr, see https://www.loc.gov/standards/iso639-2/php/code_list.php). |
| 10 | Resource Type | | M | M | Choose one of the following terms from the table |
| 11 | Alternate Identfier(s) | | O | O | Alternative identifiers (next to the one supplied in 1) uniquely describing your dataset. |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 12a | Related Item(s) | | R | R | Information about a resource related to the one being registered, e.g. another dataset based on the same source data or a publication involving this dataset . |
| 12b | | Identifier | R | R | State the persistent identifier of the related item (for example a DOI). If no persistent identifier is available use the URL. |
| 12c | | Relation Type | R | R | The particular relation to the resource should be described by one of the terms in the table |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 13 | Size | | O | O | The size (MB, GB, TB) of your dataset, in most cases the repository will calculate this for you. |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 14 | Format | | O | O | Technical formats of your data (for example pdf, xls, stata). This will help other researchers to use your data and provides information on the long term preservation of the data.Consider adding a README file to your dataset to provide a more in-depth explanation on which software you used to create your dataset. |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 15 | Version | | O | O | Version number of your dataset. Useful if you need to publish an updated version of your dataset later. |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 16 | Rights | | M | M | Provide information about how other researchers can use your dataset.If your dataset is Open, e.g. other researchers will be able to access it you should provide a license under which they can do so. For standard licenses provide a URL such as https://creativecommons.org/licenses/by-sa/4.0/If you need to use a custom license, provide it as a text file called `license.txt`. |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 17 | Description | | M | M | Describe your dataset, e.g. the subject, the sample size, methodology, etc. It is best to keep this description concise. More elaborate documentation should be added in a text file called README.If the data is meant as replication data for a publication you can reference the publication here, it is also strongly recommended to use the relation properties. |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 18 | GeoLocation(s) | | R | R | If your data is linked to a particular location provide a place name (English preferred) and/or the coordinates. Coordinates can either be a point location (as: longitude, latitude) or a bounding box defined by 4 coordinates (as: west longitude, east longitude, north latitude, south latitude) |

| ID | Property | Subproperty | Publishing | Archiving | Explanation |
|---|---|---|---|---|---|
| 19 | Funding Reference(s) | | O | O | The name(s) of the organization(s) funding the research. If using this property also add the Award Number. |

**Resource types**

| Option | Definition |
|---|---|
| **Audiovisual** | A series of visual representations imparting an impression of motion when shown in succession. May or may not include sound. |
| **Book** | A medium for recording information in the form of writing or images, typically composed of many pages bound together and protected by a cover |
| **BookChapter** | One of the main divisions of a book. |
| **Collection** | An aggregation of resources, which may encompass collections of one resourceType as well as those of mixed types. A collection is described as a group; its parts |
| **ComputationalNotebook** | A virtual notebook environment used for literate programming |
| **ConferencePaper** | Article that is written with the goal of being accepted to a conference |
| **ConferenceProceeding** | Collection of academic papers published in the context of an academic conference |
| **DataPaper** | A factual and objective publication with a focused intent to identify and describe specific data, sets of data, or data collections to facilitate discoverability |

| **Option** | Definition |
|---|---|
| **Dataset** | Data encoded in a defined structure |
| **Dissertation** | A written essay, treatise, or thesis, |
| **Event** | A non-persistent, time- based occurrence |
| **Image** | A visual representation other than text |
| **InteractiveResource** | A resource requiring interaction from the user to be understood, executed, or experienced |
| **Model** | An abstract, conceptual, graphical, mathematical or visualization model that represents empirical objects, phenomena, or physical processes |
| **OutputManagementPlan** | A formal document that outlines how research outputs are to be handled both during a research project and after the project is completed |
| **PeerReview** | Evaluation of scientific, academic, or professional work by others working in the same field |
| **PhysicalObject** | An inanimate, three- dimensional object or substance |
| **Preprint** | A version of a scholarly or scientific paper that precedes formal peer review and publication in a peer-reviewed scholarly or scientific journal |
| **Report** | A document that presents information in an organized format for a specific audience and purpose |
| **Service** | An organized system of apparatus, appliances, staff, etc., for supplying some function(s) required by end users |
| **Software** | A computer program other than a computational notebook, in either source code (text) or compiled form. Use this type for general software components supporting scholarly research. Use the "ComputationalNote book" value for virtual notebooks. |
| **Sound** | A resource primarily intended to be heard |
| **Standard** | Something established by authority, custom, or general consent as a model, example, or point of reference |

| **Option** | Definition |
|---|---|
| **Text** | A resource consisting primarily of words for reading that is not covered by any other textual |
| **Workflow** | A structured series of steps which can be executed to produce a final outcome, allowing users a means to specify and enact their work in a more reproducible manner |
| **Other** | |

**Contributor types**

| **Option** | **Definition** |
|---|---|
| **ContactPerson** | Person with knowledge of how to access, troubleshoot, or otherwise field issues related to the resource |
| **DataCollector** | Person/institution responsible for finding or gathering/collecting data under the guidelines of the author(s) or Principal Investigator (PI) |
| **DataCurator** | Person tasked with reviewing, enhancing, cleaning, or standardizing metadata and the associated data submitted for storage, use, and maintenance within a data centre or repository |
| **DataManager** | Person (or organisation with a staff of data managers, such as a data centre) responsible for maintaining the finished resource |
| **Distributor** | Institution tasked with responsibility to generate/disseminate copies of the resource in either electronic or print form |
| **Editor** | A person who oversees the details related to the publication format of the resource |
| **HostingInstitution** | Typically, the organisation allowing the resource to be available on the internet through the provision of its hardware/software/operating support |

| Option | Definition |
|---|---|
| **Producer** | Typically, a person or organisation responsible for the artistry and form of a media product |
| **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 |
| **RegistrationAgency** | Institution/organisation officially appointed by a Registration Authority to handle specific tasks within a defined area of responsibility |
| **RegistrationAuthority** | A standards-setting body from which Registration Agencies obtain official recognition and guidance |
| **RelatedPerson** | A person without a specifically defined role in the development of the resource, but who is someone the author wishes to recognize |
| **Researcher** | A person involved in analysing 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. |
| **ResearchGroup** | Typically refers to a group of individuals with 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 |
| **Sponsor** | Person or organisation that issued a contract or under the auspices of which a work has been written, printed, published, developed, etc. |

| Option | Definition |
|---|---|
| **Supervisor** | Designated administrator over one or more groups/teams working to produce a resource, or over one or more steps of a development process |
| **WorkPackageLeader** | A Work Package is a recognized data product, not all of which is included in publication. The package, instead, may include notes, discarded documents, etc. The Work Package Leader is |

**Relation types**

| Option | Definition |
|---|---|
| **IsCitedBy** | indicates that B includes A in a citation |
| **Cites** | indicates that A includes B in a citation |
| **IsSupplementTo** | indicates that A is a supplement to B |
| **IsSupplementedBy** | indicates that B is a supplement to A |
| **IsContinuedBy** | indicates A is continued by the work B |
| **Continues** | indicates A is a continuation of the work B |
| **Describes** | indicates A describes B |
| **IsDescribedBy** | indicates A is described by B |
| **HasMetadata** | indicates resource A has additional metadata B |
| **IsMetadataFor** | indicates additional metadata A for a resource B |
| **HasVersion** | indicates A has a version (B) |
| **IsVersionOf** | indicates A is a version of B |
| **IsNewVersionOf** | indicates A is a new edition of B, where the new edition has been modified or updated |
| **IsPreviousVersionOf** | indicates A is a previous edition of B |
| **IsPartOf** | indicates A is a portion of B; may be used for elements of a series |
| **HasPart** | indicates A includes the part B |
| **IsPublishedIn** | indicates A is published inside B, but is independent of other things published inside of B |
| **IsReferencedBy** | indicates A is used as a source of information by B |

| **Option** | Definition |
|---|---|
| **References** | indicates B is used as a source of information for A |
| **IsDocumentedBy** | indicates B is documentation about/ explaining A; e.g. points to software documentation |
| **Documents** | indicates A is documentation about B; e.g. points to software documentation |
| **IsCompiledBy** | indicates B is used to compile or create A |
| **Compiles** | indicates B is the result of a compile or |
| **IsVariantFormOf** | indicates A is a variant or different form of B |
| **IsOriginalFormOf** | indicates A is the original form of B |
| **IsIdenticalTo** | indicates that A is identical to B, for use when there is a need to register two separate instances of the same resource |
| **IsReviewedBy** | indicates that A is reviewed by B |
| **Reviews** | indicates that A is a review of B |
| **IsDerivedFrom** | indicates B is a source upon which A is based |
| **IsSourceOf** | indicates A is a source upon which B is based |
| **IsRequiredBy** | Indicates A is required by B |
| **Requires** | Indicates A requires B |
| **Obsoletes** | Indicates A replaces B |
| **IsObsoletedBy** | Indicates A is replaced by B |