MEMO

Guideline: Minimal Metadata for Research Data and Infrastructure

Author	Peter Vos, Alexey Pristupa, Brett Olivier
Assigned by	Stephan Okhuijsen and Marcel Ras
Date (version)	2021-11-16 (v1.0)
URL working version	(gebruik deze link voor suggesties en commentaar)
Distribution	

Goal of this memo

In this memo a recommendation is made for the establishment of a set of minimal metadata properties and their use in the description of research datasets.

Introduction

Repositories, publication platforms and other RDM-tools all have their own ways of describing the content they have to hold and ways to add metadata. Development of a set of metadata guidelines will provide researchers and support staff with an overview of the minimal metadata requirements that enables them to make their data more FAIR.

Metadata contains information about a dataset and/or its content and is an important element of the FAIR data principles. The ultimate goal of FAIR is to optimise the reuse of data. The principles of FAIR data management are specifically mentioned in the VU Research Data Management policy[1] principle 2 (3.3.25):

> *"Contribute, where appropriate, towards making data findable, accessible, interoperable and reusable in accordance with the FAIR principles."*

Making data **A**ccessible is later referenced as a researcher responsibility[2] while the VU has the responsibility to ensure data **F**indability[3]. Clearly these policy points will be facilitated by the creation of practical metadata guidelines that apply to both VU researchers and the providers of Research Data Management (RDM) infrastructure.

The VU Research Data Management policy also asks for datasets to be registered in the Pure Research Information System. Pure registrations can take place via a number of different routes (for example, sourced from different RDM systems) and a standardised approach to writing/filling-in metadata properties would facilitate this process.

This memo proposes the use of a set of minimal metadata properties that will:

1. guide the level of support for and implementation of metadata in VU RDM infrastructure,
2. act as a guideline for entering/filling-in metadata when using external RDM infrastructure.

[1] Version 2.0, 17-2-2020
[2] Responsibility 4
[3] Responsibility 5

VU VRIJE UNIVERSITEIT AMSTERDAM

To be decided

1. We propose the minimal metadata guidelines to be published as a recommendation for further operationalisation of VU RDM policy
2. We propose the status of the minimal metadata guidelines be initially "strongly recommended" (realisable in the short term) rather than a formal obligation.
3. We propose to set up an awareness campaign about the existence and use of the minimal metadata guidelines among researchers, policy makers and support staff
4. We propose that the minimal metadata guidelines should be maintained by the VU Library to ensure their implementation and future development.

Next steps

1. Finalize and publish the minimal metadata guidelines
2. Develop an awareness campaign (online information, workshops, presentations)
3. Write RDM systems user guides and maintain a VU OSF metadata document resource
4. Organise ownership and structural management

Minimal metadata guidelines

The following criteria were used in defining a minimal set of metadata properties:

1. No reinventing the wheel - if possible use existing metadata standards.
2. The set of properties should be interoperable with existing metadata standards.
3. The set of metadata properties should describe the dataset[4] as a whole and not include detailed information on the dataset contents.
4. Metadata properties should be readable/writable by both humans and machines.
5. Metadata properties should be “easy to use” for researchers and support staff.
6. Metadata properties should be general and not domain-specific

Practically, this implies that the properties of the VU-RDM minimal metadata standard should be derived from existing general metadata standards for identification, citation and retrieval of data. Some commonly used metadata standards for data include DataCite[5], CERIF[6], DCAT[7] and Dublin Core[8].

[4] A dataset is one or more research objects that are logically related and can be stored (encoded/serialised) digitally. Collections of research objects form a dataset that can be encoded, as one or more files, using standard encoding formats or in a researcher defined way.
[5] https://schema.datacite.org/meta/kernel-4.4/
[6] https://www.eurocris.org/Uploads/Web%20pages/CERIF-1.6/documentation/MInfo.html
[7] https://www.w3.org/TR/vocab-dcat-2/
[8] https://dublincore.org/specifications/dublin-core/

DataCite Metadata Schema 4.4 (2021) best satisfies these criteria in that it:

- is widely used and has a broad community acceptance,
- is interoperable with other metadata standards,
- is used by many important data repositories and
- enjoys extensive support by most RDM-tools without being overly complex.

In defining the VU-RDM minimal metadata guidelines we have followed DataCite's "three levels of obligation".

- *Mandatory* properties (6) are the minimum requirements for findability of datasets and correct registration in Pure.
- In order to enhance the prospects that metadata and datasets will be found, cited and linked to original research the use of the *Recommended* (6) and *Mandatory* set of properties is strongly encouraged.
- *Optional* properties (8) ensure a richer description of the dataset for enhanced reusability.

These proposed properties are derived from and fully compatible with DataCite. The guidelines are in some cases stricter than DataCite (for example Description is Mandatory not Recommended). For ease of use we have simplified the property names to make the guidelines less technical and easier to use. Appendix 3 shows the mapping to DataCite.

Full details of the properties and obligations of the current proposed specification are, for convenience, included as Appendix 4. The complete guidelines can be found on the vu-rdm-tech github repository[9]. An example of the filled-in minimal metadata is given in Appendix 1.

Mapping to VU infrastructure and documentation

Mapping to VU and other infrastructure

To make sure the proposed schema can actually be used in the real world we looked at the compatibility of currently used and (future) supported research systems: Pure, DataverseNL, Yoda, OSF, Figshare and Zenodo.
In all these systems it is possible to comply with at least the Mandatory properties. This is especially important for external systems like OSF and Zenodo where we have little or no influence on the used schema. Compatibility with Yoda is in our own control (Appendix 2). The minimal metadata guidelines are also compatible with the current minimum requirements of Pure. This is an important requirement for making automated registration of published datasets in Pure possible, saving time for the researcher. Appendix 3 provides compliance of supported and commonly used systems.

[9] https://github.com/vu-rdm-tech/metadata/blob/main/minimal_metadata_guide.md

For storage infrastructure that has no built-in metadata support (e.g. Research Drive, SurfDrive) the minimal metadata specification can be implemented through mapping to a formatted text file[10] or as DataCite XML.

Documentation and user manuals

To maintain compatibility of supported systems, detailed documents describing the mapping of metadata for each system should be created and maintained for the benefit of functional administrators and data stewards.

Researchers should not be burdened with interpreting the minimal metadata guidelines and should be provided with easy-to-use guidelines that describe the mapping to commonly used RDM systems. Each of these documents should use the same terminology as much as possible. Guidelines for supported and commonly used repositories (Zenodo, Figshare) should be created (Example: Yoda metadata help). In addition, guidelines for domain specific RDM systems can be developed in collaboration with Data Stewards and researchers.

Table 2. Current status of metadata mapping and user guides writing for the commonRDM systems.

RDM systems in use	VU has control	Mapping status	User guides
Yoda	Yes	Completed (see Appendix 3 for an example)	Available
Pure	Partially	Ongoing	Being drafted
DataverseNL	Partially	Needs to be initiated	Needs to be initiated
OSF	No	Ongoing	Being drafted
Zenodo	No	Needs to be initiated	Needs to be initiated
Figshare	No	Needs to be initiated	Needs to be initiated
RD/OneDrive/etc	Partially	Draft concept developed[11]	Needs to be initiated

[10] https://github.com/vu-rdm-tech/melite-metadata/blob/main/melite-proposed.md
[11] Concept text-based encoding available on GitHub VU-RDM-TECH

Appendix 1. Example of the minimal metadata properties

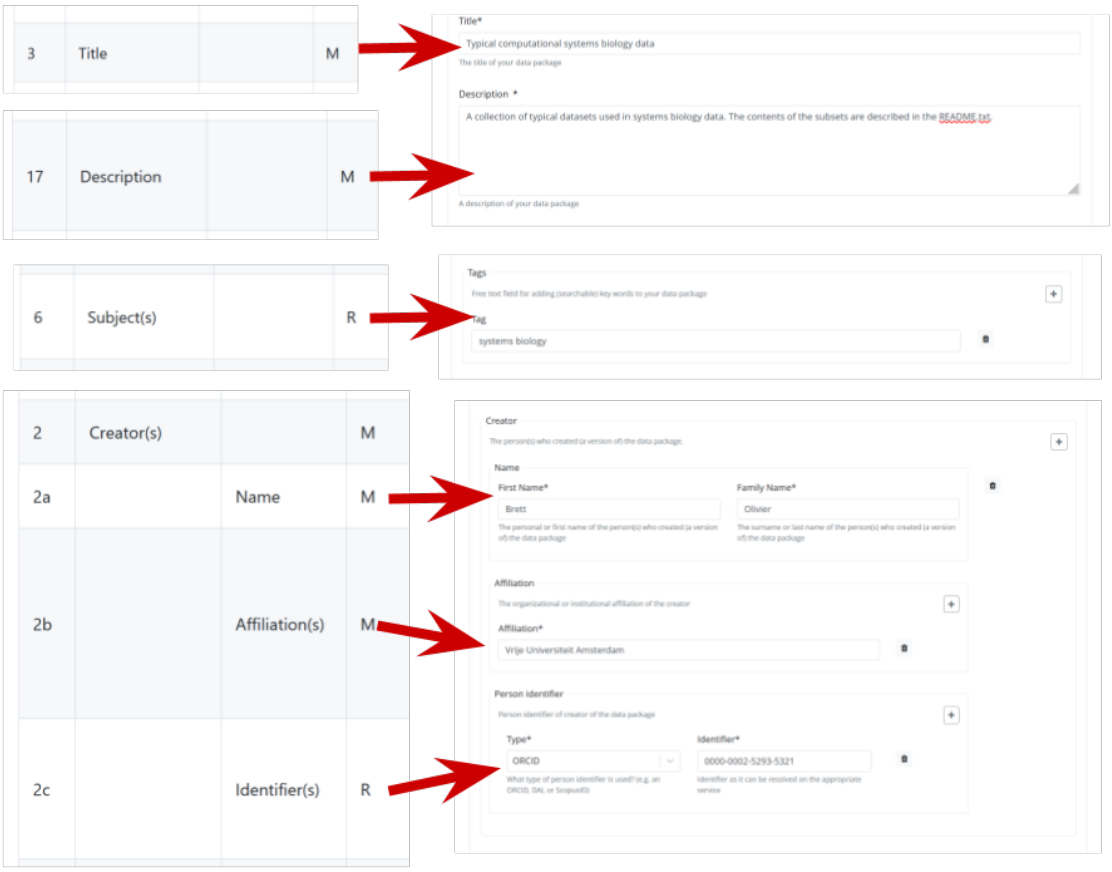
The minimal metadata associated with a hypothetical dataset:

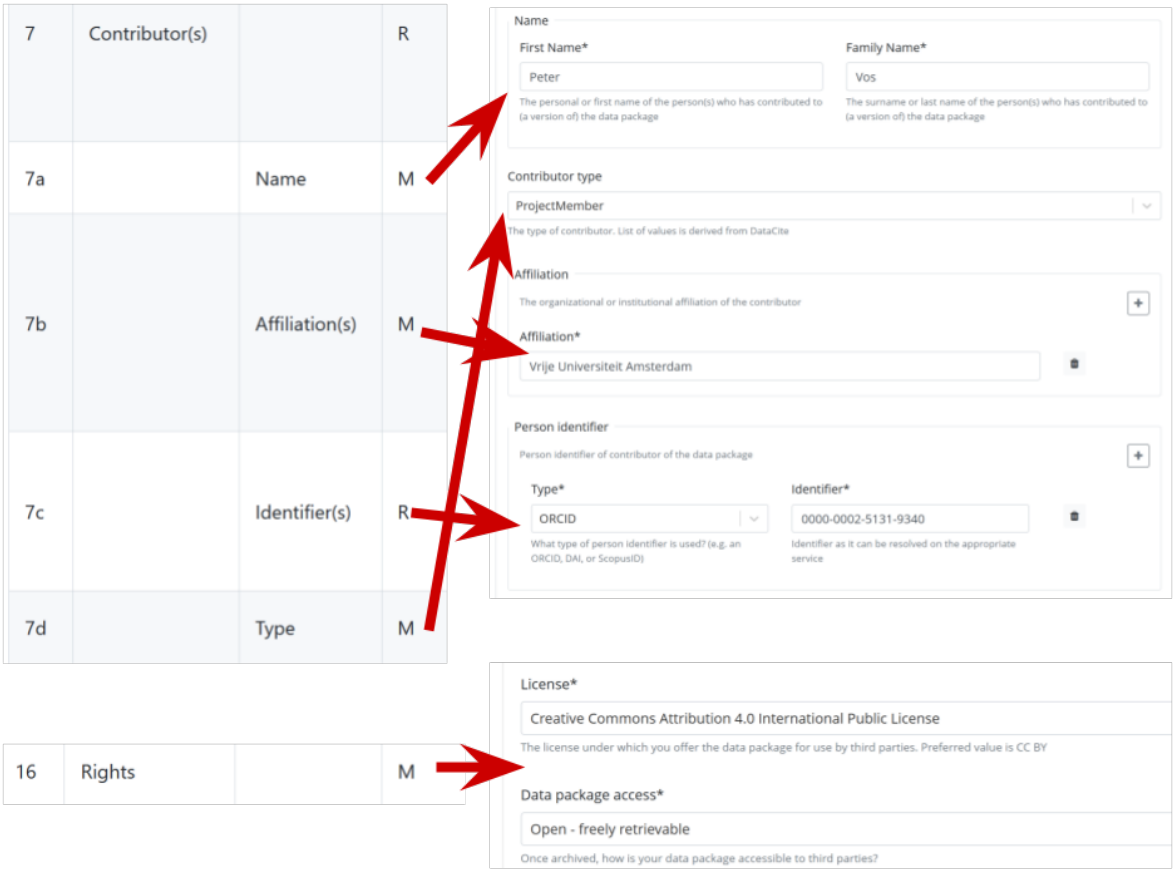
Property	Subproperty	Example value
Identifier		DOI: 10.48338/12345
Creator(s)	Name	Olivier, Brett
	Affiliation(s)	Vrije Universiteit Amsterdam AIMMS Institute
	Identifier(s)	ORCID: 0000-0002-5293-5321
Title		Typical computational systems biology data
Publisher		Vrije Universiteit Amsterdam
Publication Year		2021
Subject(s)		systems biology
Contributor(s)	Name	Vos, Peter
	Affiliation(s)	Vrije Universiteit Amsterdam
	Identifier(s)	ORCID: 0000-0002-5131-9340
	Type	ProjectMember
Date(s)		2021-10-06
Resource Type		Dataset
Rights		CC-BY-SA 4.0
Description		A collection of files containing experimental cell growth data, ODE based model descriptions, Python analysis scripts and results. The contents of the individual files and protocols used to generate them are described in the README.md.

Appendix 2. Example of mapping in Yoda

Every data repository will have subtly different property names, but in most cases these will easily map to our properties. To help researchers we provide a metadata help page for Yoda, based on the general minimal metadata Guidelines, see https://yoda.vu.nl/site/getting-started/metadata-add.html.

This is how the guidelines properties look in the Yoda metadata web form.





The somewhat more technical (sub)properties like “Contributor type” and “Person identifier type” are implemented as simple drop-down boxes. On publication, this metadata will be converted to the valid DataCite format to enable registration of the dataset on https://datacite.org.

Appendix 3: Compliance of supported and commonly used RDM systems

ID	Property	Subproperty	Publishing	Archiving	*DataCite*	*Pure*	Yoda	OSF	DataverseNL	Zenodo	Figshare
1	Identifier		M	O	M	M	A (DOI)	A (DOI)	A (DOI)	A (DOI)	A (DOI)
2a	Creator(s)	Name	M	M	M	M	M	M	M*	M	M*
2b		Affiliation(s)	M	M	O	M	M	M	R	M	O
2c		Identifier(s)	R	R	O	O	R	O	R	R	O
3	Title		M	M	M	M	M	M	M	M	M
4	Publisher		M	O	M	M	A	A	A	A	A
5	Publication Year		M	O	M	M	A	A	A	A	A
6	Subject(s)		R	R	R	-	R	R	M	O	M
7a	Contributor(s)	Name	R	R	R	O	R	O	-*	O	-*
7b		Affiliation(s)	M	M	O	M	M	O	-	M	-
7c		Identifier(s)	R	R	O	O	R	O	-	O	-
7d		Type	M	M	M	O	R	O	-	M	-
8	Date(s)		R	R	R	O	R	-	O	-	-
9	Language		O	O	O	-	O	-	O	O	-
10	Resource Type		M	M	M	-	M	M	A	M	M
11	Alternate Identifier(s)		O	O	O	-	-	-	O	-	-
12a	Related Item(s)		R	R	R	O	R	-	R	R	-
12b		Identifier	R	R	R	O	R	-	R	R	-
12c		Relation Type	R	R	R	O	R	-	R	R	-

13	Size		O	O	O	-	-	-	A	-	-
14	Format		O	O	O	-	-	-	A	-	-
15	Version		O	O	O	-	O	-	A	O	-
16	Rights		M	M	O	-	M	M	M	M	M
17	Description		M	M	R	O	M	R	M	M	M
18	GeoLocation(s)		R	R	R	O	R	-	O	-	-
19	Funding Reference(s)		O	O	O	-	O	-	O	O	O

** The DataverseNL and Figshare metadata schemas define an Author property which is used for both creators and contributors.*

The green columns show the proposed minimal properties, the yellow column shows the corresponding obligation in the DataCite schema and the blue column shows the requirements for registration in Pure. The columns to the right show the requirements of (future) supported and currently used repositories. In all these repositories it is possible to comply with the proposed minimal metadata. The minimal set is also supported by the VU CRIS Pure and complies with the minimum requirements of Pure itself.

M - Mandatory | R - Recommended | O - Optional | A - This value is automatically filled in by the repository | - Property not present

Appendix 4: 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

There is a slight difference between obligations for a published vs an archived set: Identifier (meaning a persistent public URL such as a DOI, a dataset still needs a unique identifier in the Archive itself) and publication related properties Publisher and Publication Year are optional.

ID	Property	Subproperty	Publish	Archive	Explanation
1	Identifier		M	O	This should be a global unique identifier, which preferably is unchangeable 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 afterwards. (A persistent identifier is optional for unpublished archived datasets.)
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.
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
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
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 property only applies to published datasets.)
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 property 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.
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.
7a		Name	M	M	Enter names of persons as: <family name>, <first name> <initials> e.g. Olivier, Brett G.
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
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 below for possible types. If contributor is used then Contributor Type is mandatory.
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 below
11	Alternate Identfier(s)		O	O	Alternative identifiers (next to the one supplied in 1) uniquely describing your dataset.
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 below
13	Size		O	O	The size (MB, GB, TB) of your dataset, in most cases the repository will calculate this for you.
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.
15	Version		O	O	Version number of your dataset. Useful if you need to publish an updated version of your dataset later.
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.
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.
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)

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
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
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
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.
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
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