Addendum: Minimal Metadata for Research Data and Infrastructure

In the document "Minimal Metadata for Research Data and Infrastructure" various criteria were mentioned to guide the development of the VU minimum metadata (VUMM) guidelines (here listed in Appendix 1). In order to satisfy these criteria which focus on interoperability and usability, various existing standards were considered. Two of these, the DataCite Metadata Schema and Dublin Core were noted as being well suited as potential frameworks within which to develop the VUMM guidelines. In this document we provide more information on DataCite and the DataCite Metadata Schema.

For reference purpose we provide with this document the following supporting material:

- The DataCite MetaData Specification v4.4: DataCite-MetadataKernel_v4.4.pdf
- The DataCite MetaData mapping to Dublin Core: DataCite4.4_DublinCore_Mapping.pdf

DataCite metadata: origin and organisation

The DataCite organisation is a DOI provider with the focus on research data and is analogous to the CrossRef registration infrastructure used by publishers¹.

"The global consortium DataCite was established in 2009 to overcome the challenges of data citation. The aim of the consortium is to establish easy access to data, to increase the acceptance of data publication and to support data archiving."

DataCite is global organization with an international executive board² (NL is currently represented by Dr Ingrid Dillo, DANS Deputy Director. Central to the DataCite infrastructure is the DataCite Metadata store³ (MDS) which stores the metadata associated with all registered objects (assigned a DOI). In order to register an object and receive a DOI, the metadata must be uploaded using the DataCite Metadata Schema^{1,4} maintained by the Metadata Working Group⁵ (NL is currently represented by Madeleine de Smaele, TU Delft). Indexed metadata is made available through DataCite's Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) service⁶ and is searchable using the DataCite Search tool^{1,7}

Summary. The DataCite Metadata Schema is (1) coupled to a DOI provider and (2) subscribed to by a number of international repositories and institutions⁸. We find that these are sufficient conditions to assume that it has the necessary organisational support and potential longevity as a framework for the development of the VUMM guidelines.

Datacite metadata: repository support

One of our key criteria was the ability of existing VU (and other) RDM tools to support the VUMM guidelines. To start with we considered two potential VU archiving tools Yoda and DataversNL and two external tools known to be used by researchers, Zenodo and Figshare. In Table 1, using the

¹ Neumann, J., Brase, J. *DataCite and DOI names for research data*. J Comput Aided Mol Des 28, 1035–1041 (2014). https://doi.org/10.1007/s10822-014-9776-5

² DataCite Governance, https://datacite.org/governance.html

³ DataCite Metadata Store (MDS), https://mds.datacite.org/?lang=en

⁴ DataCite Metadata Schema Repository, http://schema.datacite.org/

⁵ DataCite Metadata Working Group, https://datacite.org/steering.html

⁶ OAI-PMH https://oai.datacite.org/oai/

⁷ DataCite Search, http://search.datacite.org

⁸ DataCite members, https://datacite.org/members.html

re3data.org registry⁹ of data repositories we determined whether these tools supported either Dublin Core (DC) or the DataCite Metadata Schema (DCMS).

| | Dublin Core | DataCite Metadata Schema | |
|-------------------|-------------|--------------------------|--|
| Yoda yes | | yes | |
| DataversNL | yes | no ¹⁰ | |
| Zenodo yes | | yes | |
| Figshare yes | | yes | |

Table 1: Selected RDM tools metadata schema support

| Metadata standard | Number of data repositories | Domain |
|-------------------------------------|-----------------------------|----------------|
| Dublin Core | 425 | general |
| DataCite Metadata Schema | 254 | data citation |
| DDI - Data Documentation Initiative | 205 | social science |
| ISO 19115 | 173 | geographic |

Table 2: Re3data.org top four reported metadata standards.

Table 1 shows that from our small selection of tools all are registered as supporting Dublin Core while 75% claim DataCite Metadata Schema support. Note that from practical experience DataVerseNL does indeed export metadata as DataCite. Table 2 shows how the more generic Dublin Core is supported by the highest number of Re3data registered repositories while the more data oriented DataCite leads the more specific metadata standards.

Summary. While not as widely used as the more generic Dublin Core, the DataCite Metadata Schema is supported directly by a significant number of VU RDM tools and registered research data repositories to be considered a broadly supported community standard.

DataCite metadata: interoperability with Dublin Core

We have seen that both Dublin Core and the DataCite Metadata Schema (DCMS) are potential candidates in terms of organisational infrastructure and support by registered data repositories. In

¹⁰ Although not registered with Re3data, DataVersNL uses DataCite DOI's and can export metadata as both Dublin Core and DataCite.



⁹ Re3data registry, https://www.re3data.org/ accessed October 2021

this section we consider criteria 2, 3 and 5 (Appendix 1) with the focus on interoperability, usability and suitability for the description of datasets.

We now compare the interoperability of the DataCite Metadata Schema properties to those of Dublin Core. DataCite provides a mapping to Dublin Core shown in the supplementary document "DataCite4.4_DublinCore_Mapping.pdf"¹¹ and in this document that, in general, there is a mapping from most DataCite properties to Dublin Core. Properties not included in Dublin Core are, for example, some of the <class>Type and <class>Scheme properties (nameType, contributorType, affiliationIdentifierScheme, rightsIdentifierScheme) as well as more detailed descriptors such as givenName, familyName and Version. In Appendix 2 and 3 we expand our tables of VUMM properties to include the mapping to Dublin Core and note that the only notable non-mappable properties are contributorType and Version. Analogous to other DataCite <class>Type properties, contributorType contains a controlled vocabulary defining the role of the Contributor (see Appendix 4 for a full list of contributorTypes) and therefore includes a more accessible description of a contributors role¹².

In summary. DataCite in general and our selected VUMM properties in particular, are sufficiently interoperable with Dublin Core to take advantage of its widespread use and established status as a W3C standard.

Conclusion

In the previous sections we have shown that the DataCite Metadata Schema: has the necessary organisational support and potential longevity as a data citation standard, is supported directly and indirectly by a significant number of registered data repositories and is sufficiently interoperable with Dublin Core to take advantage of its widespread use and established status.

Through the DataCite to Dublin Core mapping we have also shown that there is a high degree of compatibility between the VUMM guideline and Dublin Core properties. However, to achieve a similarly high compatibility between the VUMM properties and DataCite Metadata Schema requires them to include the DataCite concepts of mandatory/optional/recommended property obligation and the more strictly defined dataset specific properties, such as Version and controlled <class>Types.

By ensuring DataCite Metadata interoperability we therefore enable both the FAIR description of VU data as well as leverage the advantages of being interoperable with **both** the DataCite Metadata Schema **and** Dublin Core.

¹¹ Mappings to other metadata schemas are provided in the specification appendices. For example, "DataCite-MetadataKernel_v2.1.pdf" includes mappings to the IDF Metadata Kernel, OECD Dataset and DDI schemas; "DataCite-MetadataKernel_v4.4.pdf" includes mappings to the FORCE11 Software Citation Principles. ¹² Note that the Contributor name, affiliation and identifier can still be mapped to Dublin Core irrespective of the contributorType property as DataCite 4.4 defines an "Other" type allowing for the reverse mapping.



Appendix 1: Criteria used in the development of the VUMM guidelines

- 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 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

Appendix 2: Mapping of the VU Metadata properties to Dublin Core and DataCite Metadata Schema

| ID | Property | Subproperty | Dublin Core | Publishing | Archiving | DataCite |
|-----|------------------|----------------|----------------------------|------------|-----------|----------|
| 1 | Identifier | | dcterms:identifier | М | 0 | М |
| 2a | Creator(s) | Name | dcterms:creator | М | М | М |
| 2b | | Affiliation(s) | dcterms:contributor | М | М | 0 |
| 2c | | Identifier(s) | dcterms:identifier | R | R | 0 |
| 3 | Title | | dcterms:title | М | М | М |
| 4 | Publisher | | dcterms:publisher | М | 0 | М |
| 5 | Publication Year | | dcterms:issued | М | 0 | М |
| 6 | Subject(s) | | dcterms:subject | R | R | R |
| 7a | Contributor(s) | Name | dcterms:contributor | R | R | R |
| 7b | | Affiliation(s) | dcterms:contributor | М | М | 0 |
| 7c | | Identifier(s) | dcterms:identifier | R | R | 0 |
| 7d | | Туре | Not present in Dublin Core | М | М | М |
| 8 | Date(s) | | dcterms:date | R | R | R |
| 9 | Language | | dcterms:language | 0 | 0 | 0 |
| 10 | Resource Type | | dcterms:type | М | М | М |
| | Alternate | | | | | |
| 11 | Identifier(s) | | dcterms:identifier | 0 | 0 | 0 |
| 12a | Related Item(s) | | | R | R | R |
| 12b | | Identifier | dcterms:relation | R | R | R |
| 12c | | Relation Type | All types map to dcterms | R | R | R |
| 13 | Size | | dcterms:extent | 0 | 0 | 0 |
| 14 | Format | | dcterms:format | 0 | 0 | 0 |
| 15 | Version | | Not present in Dublin Core | 0 | 0 | 0 |
| 16 | Rights | | dcterms:rights | М | М | 0 |
| 17 | Description | | dcterms:description | М | М | R |
| 18 | GeoLocation(s) | | dcterms:spatial | R | R | R |
| 19 | Funding | | dcterms:contributor | 0 | 0 | 0 |



| Reference(| s) | | |
|------------|----|--|--|

The green columns show the proposed minimal properties, the yellow column shows the corresponding obligation in the DataCite schema. M - Mandatory | R - Recommended | O - Optional

Appendix 3. Example of the minimal metadata properties

The minimal metadata associated with a hypothetical dataset. The Property and subproperty columns represent the VU minimal terms expressed in DataCite Metadata Schema 4.4. The Dublin Core column is the same (sub)property mapped to Dublin Core. In this example, only one property DataCite::ContributorType is not mapped.

| Property | Subproperty | Dublin Core | Example value | |
|---------------------|----------------|----------------------------|---|--|
| Identifier | | dcterms:identifier | DOI: 10.48338/12345 | |
| Creator(s) | Name | dcterms:creator | Olivier, Brett | |
| Affiliation(s) | | dcterms:contributor | Vrije Universiteit Amsterdam AIMMS Institute | |
| | Identifier(s) | dcterms:identifier | ORCID: 0000-0002-5293-5321 | |
| Title | | dcterms:title | Typical computational systems biology data | |
| Publisher | | dcterms:publisher | Vrije Universiteit Amsterdam | |
| Publication Year | | dcterms:issued | 2021 | |
| Subject(s) | | dcterms:subject | systems biology | |
| Contributor(s) | Name | dcterms:contributor | Vos, Peter | |
| | Affiliation(s) | dcterms:contributor | Vrije Universiteit Amsterdam | |
| | Identifier(s) | dcterms:identifier | ORCID: 0000-0002-5131-9340 | |
| | Туре | Not present in Dublin Core | ProjectMember | |
| Date(s) | | dcterms:date | 2021-10-06 | |
| Resource Type | | dcterms:type | Dataset | |
| Rights | | dcterms:rights | CC-BY-SA 4.0 | |
| Description | | dcterms: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 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 | | М | 0 | This should be a global unique identifier, which prefethe datasets. In most cases the repository where you this in the form of a Handle or DOI. If no persistent URL is available you could use a norm should not move the data afterwards. (A persistent identifier is optional for unpublished ar |
| 2 | Creator(s) | | М | М | The main researchers involved in producing the data |
| 2a | | Name | М | М | Enter names of persons as: <family name="">, <first nar<="" td=""></first></family> |
| 2b | | Affiliation(s) | М | М | Always make sure to always enter your affiliations who dataset. Make sure you at least add the VU, the correct name Note: some repositories may allow you to enter a RC Registry). The VU ROR is: https://ror.org/008xxew50 |
| 2c | | Identifier(s) | R | R | If known, enter one or more unique identifiers like A ResearcherID. The VU strongly recommends registering for an ORCI easy way to uniquely identify yourself over which you |
| 3 | Title | | М | М | A descriptive title for your dataset, should not be lon |
| 4 | Publisher | | М | 0 | Name of the organization where you published your repository where you upload your data will fill this in name of the organization owning the website or data (This property only applies to published datasets.) |
| 5 | Publication Year | | М | 0 | The year (or date) you first published your dataset. the 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. The dataset on the internet. Some repositories will have controlled term lists to controlled. |
| 7 | Contributor(s) | | R | R | The institution or person responsible for collecting, rotherwise contributing to the development of the re For software, if there is an alternate entity that "hold distributes, releases, issues, orproduces" the code, u "hostingInstitution" for the code repository. |
| 7a | | Name | М | М | Enter names of persons as: <family name="">, <first nar<="" td=""></first></family> |
| 7b | | Affiliation(s) | M | М | Always make sure to always enter your affiliations who dataset. Make sure you at least add the VU, the correct name Amsterdam". Note: some repositories may allow you to enter a RC Registry). The VU ROR is: https://ror.org/008xxew50 |
| 7c | | Identifier(s) | R | R | If known, enter one or more unique identifiers like A ResearcherID. The VU strongly recommends registering for an ORCI easy way to uniquely identify yourself over which you |
| 7d | | Туре | М | М | The role of the contributor, see the below for possibl Contributor Type is mandatory. |
| 8 | Date(s) | | R | R | If applicable add extra dates applying to your dataset collected", meaning the date or date range when you |
| 9 | Language | | 0 | 0 | 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 |
| 10 | Resource Type | | М | М | Choose one of the following terms from the table be |
| 11 | Alternate Identfier(s) | | 0 | 0 | Alternative identifiers (next to the one supplied in 1) |
| 12a | Related Item(s) | | R | R | Information about a resource related to the one beir based on the same source data or a publication invol |
| 12b | | Identifier | R | R | State the persistent identifier of the related item (for identifier is available use the URL. |
| 12c | | Relation Type | R | R | The particular relation to the resource should be des table below |
| 13 | Size | | 0 | 0 | The size (MB, GB, TB) of your dataset, in most cases t |



| | | | | you. |
|----|-------------------------|---|---|--|
| 14 | Format | 0 | 0 | Technical formats of your data (for example pdf, xls, stresearchers to use your data and provides information the data. Consider adding a README file to your dataset to propose on which software you used to create your dataset. |
| 15 | Version | 0 | 0 | Version number of your dataset. Useful if you need t your dataset later. |
| 16 | Rights | М | М | Provide information about how other researchers call f your dataset is Open, e.g. other researchers will be provide a license under which they can do so. For states https://creativecommons.org/licenses/by-sa/4.0/ If you need to use a custom license, provide it as a termination of the provide it as a ter |
| 17 | Description | М | М | Describe your dataset, e.g. the subject, the sample si keep this description concise. More elaborate docum file called README. |
| 18 | GeoLocation(s) | R | R | If your data is linked to a particular location provide a and/or the coordinates. Coordinates can either be a latitude) or a bounding box defined by 4 coordinates longitude, north latitude, south latitude) |
| 19 | Funding Reference(s) | О | О | The name(s) of the organization(s) funding the resea the Award Number. |

Resource types

| Option | Definition |
|-----------------------|--|
| Audiovisual | A series of visual representations imparting an impression of motion when shown in sound. |
| Book | A medium for recording information in the form of writing or images, typically compound 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 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 spectollections 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 experie |
| Model | An abstract, conceptual, graphical, mathematical or visualization model that represent or physical processes |
| OutputManagementPlan | A formal document that outlines how research outputs are to be handled both during project is completed |
| PeerReview | Evaluation of scientific, academic, or professional work by others working in the same |
| PhysicalObject | An inanimate, three- dimensional object or substance |
| Preprint | A version of a scholarly or scientific paper that precedes formal peer review and publi scholarly or scientific journal |
| Report | A document that presents information in an organized format for a specific audience |
| Service | An organized system of apparatus, appliances, staff, etc., for supplying some function |
| Software | A computer program other than a computational notebook, in either source code (text) for general software components supporting scholarly research. Use the "Computation notebooks." |
| Sound | A resource primarily intended to be heard |
| Standard | Something established by authority, custom, or general consent as a model, example, |
| Text | A resource consisting primarily of words for reading that is not covered by any other t |
| Workflow | A structured series of steps which can be executed to produce a final outcome, allowing enact their work in a more reproducible manner |
| Other | |

Contributor types

| | Option | Definition |
|-----|--------|------------|
| - 1 | Option | Definition |



| | i |
|-----------------------|--|
| ContactPerson | Person with knowledge of how to access, troubleshoot, or otherwise field issues related |
| DataCollector | Person/institution responsible for finding or gathering/collecting data under the guidelir Investigator (PI) |
| DataCurator | Person tasked with reviewing, enhancing, cleaning, or standardizing metadata and the asstorage, 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 resource |
| Distributor | Institution tasked with responsibility to generate/disseminate copies of the resource in e |
| 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 hardware/software/operating support |
| Producer | Typically, a person or organisation responsible for the artistry and form of a media produ |
| ProjectLeader | Person officially designated as head of project team or sub- project team instrumental in development of the resource |
| ProjectManager | Person officially designated as manager of a project. Project may consist of one or many |
| ProjectMember | Person on the membership list of a designated project/project team |
| RegistrationAgency | Institution/organisation officially appointed by a Registration Authority to handle specific responsibility |
| RegistrationAuthority | A standards-setting body from which Registration Agencies obtain official recognition an |
| RelatedPerson | A person without a specifically defined role in the development of the resource, but who recognize |
| Researcher | A person involved in analysing data or the results of an experiment or formal study. May one of the authors who helped with research but who was not so "key" as to be listed as |
| ResearchGroup | Typically refers to a group of individuals with a lab, department, or division that has a spo |
| RightsHolder | Person or institution owning or managing property rights, including intellectual property |
| Sponsor | Person or organisation that issued a contract or under the auspices of which a work has developed, etc. |
| Supervisor | Designated administrator over one or more groups/teams working to produce a resource |



| | development process |
|-------------------|---|
| WorkPackageLeader | A Work Package is a recognized data product, not all of which is included in publication. I 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 |

