# GeRDI Metadata Schema Documentation

# For Long Tail Research Data

---

GeRDI Metadata Schema

Version 2.0

January 2020

Author: Fidan Limani

Contact

https://www.gerdi-project.eu/

# Table of Contents

# Introduction

The Generic Research Data Infrastructure project (GeRDI)[1] provides an infrastructure that targets long tail research data (RD), often excluded from the infrastructure-related projects. Metadata (rather than data) are the resources or artifacts GeRDI operates on. The pilot communities, providing feedback from the beginning of the project, provided both use cases and metadata collections required to support those use cases, which, in turn, were mapped to requirements implemented incrementally, in several releases of the project. One of the key components – and deliverables – of the project is the GeRDI Metadata Schema (GeRDI Schema). This schema is used to describe or structure harvested metadata, as well as support GeRDI services, including Search, Bookmark, Store, Process, Analyze, and Submit[2].

This document focuses on documenting the different aspects that were considered – and which affected – during the schema design in GeRDI. Starting with the goals and objectives of GeRDI Schema, we present the communities involved, which will shed light on the diversities that the design process had to consider. The design-related section of the document presents the uses cases that guided the identification of services to support researchers from the pilot communities in the different phases of research. Afterwards, we present the metadata properties of the schema, and conclude with the reserved values and controlled vocabularies used for some of the properties, part of the Appendix.

## GeRDI Project: Goals and objectives

RD infrastructure projects that focus on specific domain or certain type of RD – such as Big Data – are becoming quite present. On the other hand, efforts that support long tail RD in an infrastructure context are less represented. With the goals to contribute to closing this gap, GeRDI embarked on a 3-year project that aimed to provide a set of relevant services able to support different research activities based on metadata for long tail research.

## Pilot research communities

Communities that contribute to the so-called long tail research, are typically characterized by lack of structure, i.e., no adherence to or convergence towards RD management standards, generation of RD that is small in volume (as opposed to Big Data, for example) and highly diverse in type (a lot of smaller-scoped research disciplinary RD, for example).

GeRDI pilot communities provided continuous feedback about their use cases and requirements from a RD infrastructure. With each new release in GeRDI, they evaluated the services and accompanying aspects of interest to them, including metadata. In this way, they were instrumental in both identifying and evaluating infrastructure services in this project.

The list below shows the GeRDI pilot communities. In some cases, there are multiple communities associated with a single (sub) domain, such is the case with the EREE communities, but we kept only the main communities for brevity:

- Alpine Environmental Data Analysis Center (AlpEnDAC)
- Microscopy and Bioinformatics (CBG)
- Digital Humanities (CRANE)

[1] https://www.gerdi-project.eu/.
[2] Check out the prototype: https://www.demo.gerdi.org/.

- Hydrology and River Basin Management (HFM)
- National Center for Tumor Diseases (NCT)
- Socio-Economic Panel (SOEP - DIW)
- UN International Strategy for Disaster Reduction and Environmental Computing (UNISDR)
- Digital geo-linguistics (Verba Alpina)
- Environmental, Resource and Ecological Economics (EREE)

One important aspect to mention at this point is the way to model or plan for the different research activities in these communities. As with other aspects of the project, established and structured models are always best to model and understand the requirements of the pilot communities. In GeRDI we relied on the RD lifecycle[3] from UK Data Archive to mold – put in specific research phases – the different research activities from the community members.

# GeRDI Metadata Schema: Design overview

## Use cases

For the requirements-gathering process of the project we relied on behavior-driven development process as a more agile methodology that fitted its dynamics and communities. Each community, in turn, provided their use cases, which we then matched to the (set of) service in GeRDI that was required to support. Table 1 below provides only few examples of community use cases and their mapping to the services required to implement them in GeRDI.

Table 1 Part of community use case and services mapping

| Use case | Community | Short description | Repository | Harvest | Search | Bookmark | Store | Analyze | Submit |
|---|---|---|---|---|---|---|---|---|---|
| ENA | Genetics | Visualize quality metrics (confidence level) of DNA data | Yes | Yes | Yes<br><br>Search reference genome sequences | Yes<br><br>Add to collection | Yes<br><br>Store reference genome data from ENA | Yes<br><br>Calculate quality metrics for experiment data (DNS library; short & long read data) | Yes |
| OGLP | Digital humanities | Open Greek and Latin | No | Yes | Yes<br><br>Search for ancient Greek & Latin texts | Yes | No | Yes<br><br>Topic modeling. | Yes |
| Clinical Trials | Life Sciences | Structured review of clinical studies | Yes | Yes | Yes<br><br>Search clinical trials data | Yes<br><br>Aggregated metadata to support review | Yes | Yes<br><br>Map similar terms based on text model. | Yes |

[3] GeRDI is modeled based on the UK Research Data Lifecycle: https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | process | | | |
| … | … | … | … | … | … | … | … | … | … |

As a result of the different RD lifecycles across research communities, we wanted to design GeRDI in a more modular way, with a set of services that could be selected from the different communities. The table above also represents this aspect, where it is clear that some of use cases (and potentially communities) need certain services, and do not have to necessarily rely on/deploy the complete set of GeRDI services for their research activities. The same rationale would especially apply in the wake of GeRDI adoption/use by new communities.

## Design overview

As the requirements-gathering process was advancing, so was the GeRDI Schema. After initially providing support on the general metadata aspects, we identified and introduced metadata elements required at the infrastructure level, as well as, during the later phases of the project, identify and provide support for the disciplinary metadata of interest to the communities. As a result of this design decision, although GeRDI schema changed through several of its design iterations, it is noticeable that there are roughly a **core** part, aimed to support generic metadata elements, a small extension to support the infrastructure needs, such as information about resources metadata being harvested, and a more **disciplinary** part of the schema designed to support the metadata requirements of the specific communities. Figure 1 provides the conceptual view of this design.

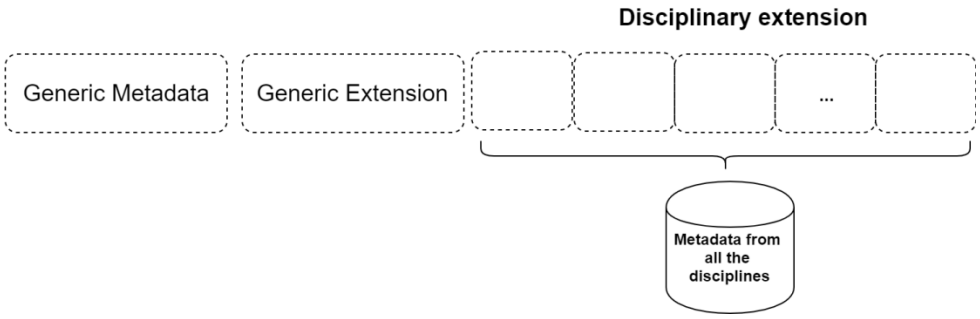


Figure 1 GeRDI Schema components

Although we have worked through several schema versions, here we present the latest stable version – version 2.0, currently implemented in GeRDI.

### *GeRDI Schema Components*

Once the conceptual design was set, we needed to provide concrete metadata solutions to each GeRDI Schema component. Established metadata standards and the availability of resources for the generic metadata requirements enabled us to address the first metadata component of the schema (considering Figure 1, we started with solutions for the components from left to right). This also proved a task with a minimal-to-no community feedback, since there were not many differences for generic metadata needs across communities. The next component we focused on, almost in parallel to that for generic metadata, was an extension that is typically required for an infrastructure that brings resources from different sources into a single place. Finally, we turned out attention to the disciplinary extension, which targeted specific metadata needs across communities. This last component of GeRDI Schema took the most time, and it lasted for the longest time of the project. Due to lack of established research (and metadata) practices with the long tail research communities, this process involved the community extensively.

### *GeRDI Schema Specification*

With the conceptual part complete, we moved to identify concrete metadata solutions for the schema, with reusability being our driving motivation throughout the process. The first two parts of the schema were easier to tackle. In this way, due to its established status and the community backing, we chose DataCite Schema[4] for the generic part of GeRDI schema. DataCite offers support for the generic metadata requirements of the pilot communities.

The "Generic extension" of the schema includes GeRDI requirements from the infrastructure point of view. There are no metadata standards that focus on this particular niche (infrastructure metadata), so we had to introduce a set of elements to concretize this component. In the process, we tried to keep the number of metadata elements that we introduce minimal (see Figure 2 for a brief list of metadata elements). Examples for this component of the schema include metadata such as the identification/link of the repository from where a metadata collection was harvested, the link (URI) of the research data location, research discipline, details about the research data themselves (file type, size, format, etc.), and so on. The "GeRDI Metadata Properties" provide details about all the metadata elements specified for this schema component.

The "Disciplinary metadata" component contains all the metadata elements that are so specific to certain communities that they cannot be represented in the "Generic metadata" part. Such examples can be seen in Figure 2. In addition, as seen in it, we organize the metadata in this section based on the research disciplines, as specified by the top-most categories from the DVG classification[5].

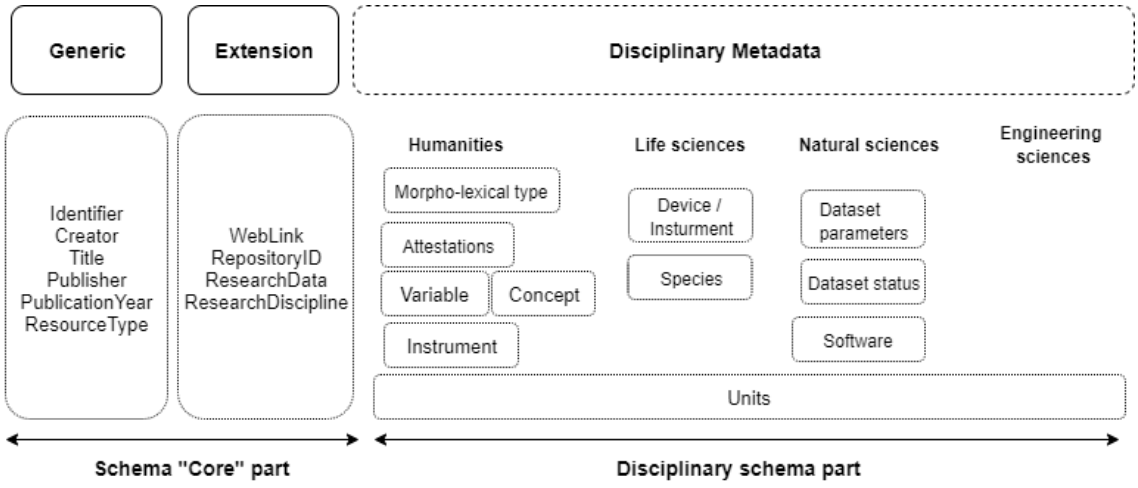


Figure 2 GeRDI Schema Specification

GeRDI schema has a stable metadata part – core part – that does not change, and a more dynamic metadata part – disciplinary part – meant to evolve as new metadata elements are provided by the communities, shown in Figure 2.

# GeRDI Metadata Properties

[4] https://schema.datacite.org/
[5] https://www.dfg.de/en/dfg_profile/statutory_bodies/review_boards/subject_areas/index.jsp

This section documents the metadata properties that form GeRDI Schema, starting with a property overview, and then moving on to a more detailed documentation about each element, including their designation for **M**andatory, **R**ecommended, or **O**ptional.

## Overview of Properties

Table 2 provides a detailed description of the GeRDI schema properties, including the designation, cardinality (Occurrence), a brief description, as well as any value range and potential restrictions. While most of the column headings from this table are self-explanatory, a brief explanation for few of them follows.

The *Designation* specifies if a metadata property's usage is a must (Mandatory), good to have (Recommended), or can be left out (Optional). The designations are important to consider as they affect the schema validation process. Namely, in case a metadata record harvested in GeRDI does not provide a value for a mandatory metadata property, it will not be stored and added to GeRDI index. The *recommended* designation is less strict, of course, for which you should try to provide a value as GeRDI services will benefit from. Due to the disciplinary variety in GeRDI, the *optional* designation enables submitting (valid) metadata records with missing values for these properties.

*Occurrence* specifies how many times a metadata record can use a property, with (1) indicating exactly one time, (0-1) indicating an optional occurrence (it can either used or not), (0-n) specifies a range from optional (not using it) to multiple times, and, in a similar case, (1-n) denotes a minimum occurrence of 1, up to multiple times.

On a final note, property names start with a capital letter, whereas sub-properties are written in a camel-case notation.

## Metadata Properties

In this section we present the metadata properties based on each component of GeRDI Schema, including the generic, infrastructure, and disciplinary parts. Due to our design decisions, the generic part of the schema is rather a pointer to the corresponding adopted standard, whereas for the remaining two components we provide the complete information based to the description from the previous section.

**Schema core: Generic Metadata**

As we already mentioned, we chose DataCite Schema for the generic part of the schema. We started with DataCite v4.1 and, as it was updated, we followed up with these updates. In the current GeRDI Schema, we rely on DataCite v4.3.

DataCite schema is well-documented, so we will not provide anything more than what has already been shown in Figure 2 (Schema "core" part) for it: there you see only the mandatory properties, without sub-properties or other details. In case you are interested to know more about, feel free to follow its documentation[6].

**Schema core: Infrastructure-specific Metadata**

The sub-section on *Design overview* already introduced the goals of the infrastructure-related metadata, as well as provided examples that fit the scenario in GeRDI. Table 2 provides a complete list of this metadata group. The property identifier is preceded with the letter "E" as it denotes the property as part of the infrastructure e*xtension* part of the schema.

[6] https://schema.datacite.org/meta/kernel-4.3/doc/DataCite-MetadataKernel_v4.3.pdf

Table 2 Infrastructure Metadata Properties

| ID | Property | Designation | Occ | Definition | Allowed values, examples, other constraints |
|---|---|---|---|---|---|
| E1 | WebLink | R | 0-n | A string that identifies a resource in GeRDI, via a (name, URI) pair. | The name is a string, not constrained in any way, whereas linkURI should be a valid URI. |
| E1.1 | webLinkName | O | 0-1 | String value denoting the name of the link (sth that the user would see during browsing) | e.g.: FishBalticSea |
| E1.2 | webLinkURI | M | 1 | The (access) URI of the resource | |
| E1.3 | webLinkType | R | 0-1 | The type of the weblink. | ViewURL, SourceURL, ProviderLogoURL, ThumbnailURL, Related |
| E2 | RepositoryIdentifier | M | 1 | A unique human readable string that identifies the source repository. | |
| E3 | ResearchData | M | 1-n | A downloadable file from the source repository. | |
| E3.1 | researchDataIdentifier | M | 1 | A universal unique identifier for the file. | |
| E3.2 | researchDataURL | M | 1 | | |
| E3.3 | researchDataLabel | R | 0-1 | A human readable name of the file | |
| E3.4 | researchDataType | R | 0-1 | The file type of the research data. | |
| E4 | ResearchDiscipline | R | 1-n | The research discipline of the data set. | E4 and its sub-types can be filled based on the DFG subject areas. |
| E4.1 | discipline | R | 1 | | |
| E4.2 | area | R | 0-1 | | |
| E4.3 | category | R | 0-1 | | |
| E4.4 | rnbr | R | 0-1 | | |

Without getting into too many details and covering all the cases, the designations for this group of metadata properties mirror their importance to the infrastructure. For example, while the *webLinkURI* sub-property is important in order to show where we retrieved certain collection from, its type, *webLinkType*, although important, can be omitted.

**Disciplinary schema part**

In identifying disciplinary metadata requirements, community involvement is a key. This process typically stretches over longer time periods, and in our case, it took several project releases before we were able to find a sufficiently-mature set of disciplinary metadata for all the pilot communities. In its final version, this part of the schema consists of 9 elements.

In order to maintain it and reuse it easier, we decided to adopt a classification for the research disciplines. This provides any research community interested in GeRDI Schema to locate, contribute, or otherwise use it. We use these top research discipline categories to group the community-specific metadata that are part of GeRDI Schema (see Figure 2).

Since disciplinary metadata are specific to certain disciplines and their designation is *optional*, we remove this column from Table 3.

Table 3 Disciplinary metadata properties

| ID | Property | Discipline | Occ | Definition |
|---|---|---|---|---|
| D1 | Variable | Humanities | 0-1 | Longitudinal study variables. |
| D1.1 | variableName | | 0-1 | |
| D1.2 | source | | 0-1 | |
| D1.3 | Concepts | | 0-n | |
| D2 | Concept | Humanities | 0-1 | Concepts for the previous variables. |
| D2.1 | conceptName | | 1 | |
| D2.2 | label | | 1 | |
| D2.3 | language | | 1 | The language the label is written in. |
| D3 | Morpho-lexical type | Humanities | 0-1 | |
| D4 | Attestation | Humanities | 0-1 | The number of speakers of a certain morpho-lexical type |
| D5 | Device/Instrument | Life sciences | 0-1 | A device/instrument used to generate the images from the microscopy collection. |
| D5.1 | deviceType | | | |
| D5.2 | deviceDescription | | | |
| D5.3 | deviceIdentification | | | |
| D6 | Species | Life sciences | 0-1 | |
| D6.1 | speciesName | | 1 | |
| D6.2 | speciesType | | 1 | |
| D7 | Dataset parameters | Natural sciences | 0-n | |
| D7.1 | parameterName | | 1 | |
| D7.2 | parameterDescription | | 1 | |
| D7.3 | parameterUnit | | 0-1 | |
| D8 | Dataset status | Natural sciences | 0-1 | |
| D8.1 | status | | 1 | |
| D8.2 | statusDescrition | | 1 | |
| D9 | Software | Natural sciences | 0-1 | |
| D9.1 | softwareDescription | | 0-1 | |
| D9.2 | softwareType | | 0-1 | |
| D9.3 | softwareVersion | | 0-1 | |

## Schema Specification

The schema specification (in YAML) is contained in the file named "metadata-index-settings.yml", and it is available on GitHub as part of the "Elasticsearch-mapping_Search" repository. You can find the complete code base for GeRDI, structured in over 50 repositories at https://github.com/GeRDI-Project.

# Appendix

**Attribute Values and Controlled Vocabularies**

In this section we provide the allowed attribute values as well as the controlled vocabularies of the GeRDI Schema metadata properties for which such restrictions exist.

E1.3 webLinkType

| Value | Definition |
|---|---|
| ViewURL | URI used to view a resource (metadata) representation. |

| | |
|---|---|
| SourceURL | URI that gives us access to the underlying data of a resources. |
| ProviderLogoURL | URI logo for the resource |
| ThumbnailURL | Thumbnail URI of the resource |
| Related | |

E4 ResearchDiscipline

| **Value** | **Definition** |
|---|---|
| DFG subject area classification | We already introduced the link to this resource. |