Metadata and documentation

Metadata is any information about data, i.e. **descriptors** that make it easier to classify and find them. Documenting all information necessary to make research data intelligible to others (i.e. anyone not involved in its collection, anyone who wants to reuse it, or even your future self) is crucial to proper research data management. Metadata is a special form of documentation: by presenting information in a **structured** way, they describe, explain and locate data, and facilitate their use. Metadata are **machine-readable** and, as such, make data findable by search engines, allow software/services to access, understand, transform, export and/or import the data to/from other software.

Types of metadata

There are different types of metadata, including:

- **Structural metadata** provide information about how the object (e.g. file, dataset, etc.) is organised. For example, structural metadata may describe the layout of a table or the relationships between its elements.
- **Descriptive metadata** provide information about the content of an object (e.g. file, dataset, etc.) and describe it in an essential manner. For example, the place and the time a photograph was taken are descriptive metadata.
- **Data quality metadata** collect specific elements to define indicators of the quality of an object. It may include quantitative metrics, such as the variance or standard error of a variable, or more qualitative aspects (e.g. a discrete classification).
- **Administrative metadata** include elements such as object identifiers (e.g. the Persistent Identifier associated with a dataset). Administrative metadata also include:
  - **Provenance metadata**, documenting information about the authors and processes that produced the object. This category includes the names and affiliations of the authors.
  - **Legal metadata**, documenting the conditions for (re)using the data. This category includes information about licences.

TITLE: Panorama of Bologna

LOCATION: Bologna

DATE: 01/01/2024

EXPOSURE: 1/30s

Image credits: Photo by Bianca Ackermann for Unsplash

Standard metadata schemas

Metadata are usually structured according to schemas, meaning their elements are organised based on **standardised conventions**. Using a standard offers many advantages in terms of interoperability, making it possible to **assign a unique, unambiguous meaning to each element**, gather essential information about the data, and compare heterogeneous data from several sources, domains and disciplines.

Standard metadata schemas can be either **generic or disciplinary**.

Generic metadata schemas such as Dublin Core collect a minimum set of high-level standard information, are easy to use and widely adopted. However, when more specific information is required, they often need to be expanded.

Domain-specific schemas have a much richer vocabulary and structure, tend to be highly specialised, and tend to be fully understandable only by researchers in that domain.

A further level of standardisation can involve the **vocabularies** employed to fill in the values of the metadata elements **Using different types of standards**.

In the field!

I am a researcher, and I am looking for a disciplinary metadata schema to describe my data. *Where do I start?*

Regardless of your research domain, there are several online portals you can use to look for the metadata schemas that best suit your needs:

1) FAIRsharing, a manually curated resource that gathers data and metadata standards;
2) The Research Data Alliance (RDA) Metadata Standards Directory;
3) The Digital Curation Centre (DCC) List of Metadata Standards, which gathers disciplinary metadata standards.

I am a biomedical researcher, and I am unable to find a metadata schema suitable for my data. *What can I do?*

If you are having trouble finding a specific metadata schema, you should look for the 'minimum information about your topic' (MIAME, MIAPE or MIAPPE) recommended for your specific data type in your specific discipline.

Metadata and repositories

The **repository you choose** to deposit your datasets influences the choice of metadata schema. At the time of deposit, a set of contextual information is required; this is organised according to a standard schema, and it effectively constitutes the metadata of the object deposited **Repositories**.

So, knowing from the early phases of your project what repository you will use gives you a clear idea of what metadata fields you will need to fill in and guides you in adding the right metadata to link your data to other software and systems.

In the field!

I have chosen the repository where I will deposit my data. *How do I find out which metadata schema it adopts?*

This information can be easily found in the repository documentation pages or via the re3data registry. AMS Acta, the University repository, uses the Dublin Core and DataCite schemas, just like Zenodo.

The repository I have chosen to deposit my disciplinary data uses a generic metadata standard.
How can I add more information to my data?

Even though a repository uses a generic schema, you can still add domain-specific metadata to your documentation. If you need to find a suitable schema, see the first point in this section.

Metadata and documentation

While metadata is a special type of documentation, structured and designed so that machines can read and interpret it, other kinds of documentation take the form of **human-readable text documents**, such as README files or codebooks. They gather additional information needed to understand and interpret the data they refer to. As a rule, while a README file contains high-level information, such as the scope and context of research, funding, methodologies. A codebook is designed to document, for example, the meaning of the names of the variables, or the units of measurement, if any.

A README file designed to accompany a dataset must include:

- General information about the project, such as title and goals of the project and dataset, names/roles/contact details of the researchers involved, funding information, Persistent Identifier (PID), etc.
- Folder and file structure and naming system, file names and folder structure, relationships and dependencies between files, description of the content of each main file, etc.
- Information about methods and software used for data collection (including references, documentation, links, experimental conditions, instrument standards and calibration, etc.), methods and software used for data processing, file formats and description of the version control system, quality control procedures applied, etc.

- Information about reuse and links to other materials, i.e. information about licences and restrictions on (parts of) the dataset, links to publications based on the dataset, relationships with other datasets and other resources used as a source for data collection (books, articles, etc.).

The codebook, which may also form part of the README file, is usually designed to document the meaning of the names of the variables and of the units of measurement, if any. At data collection level, it is usually a tabular file that contains:

- Definition of codes, symbols and abbreviations used in the files.
- List of variables with their full name and definition.
- Definition of column headings and row labels for tabular data.
- Units of measurement and data formats (e.g. YYYYMMDD).
- Processing of missing data (code, etc.).

Il file contenente la documentazione che descrive il dataset deve essere archiviato insieme ai dati nel momento in cui questi vengono depositati in un repository. È opportuno che il file di documentazione sia salvato in un formato aperto e accessibile (es .rtf).

Useful links

FAIR and the Notion of Metadata https://faircookbook.elixir-europe.org/content/recipes/introduction/metadata-fair.html

The Research Data Management Toolkit, Documentation and Metadata https://rdmkit.elixir-europe.org/metadata_management

The Turing Way Guide for Reproducible Research https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-metadata

FAIRify Your Data: Data Documentation and Metadata, Flora D'Anna https://osf.io/wbr7t

Research Data Management: Metadata (University College Dublin Library) https://libguides.ucd.ie/data/metadata

Dublin Core Metadata Standard https://www.dublincore.org/specifications/dublin-core/dces/

Resources to search for metadata schemas:

- FAIRsharing Standard Registry https://fairsharing.org/search?fairsharingRegistry=Standard
- RDA Metadata Standards Directory https://rd-alliance.github.io/metadata-directory/standards/
- DCC Guidance on Disciplinary Metadata https://www.dcc.ac.uk/guidance/standards/metadata

Minimum Information Standards:

- MIAME https://www.fged.org/projects/miame
- MIAPE https://www.psidev.info/miape
- MIAPPE https://www.miappe.org/