UNIVERSITÀ DEGLI STUDI DI MILANO

Dataverse quality check

Document Data

Abstract	*Proposal for a quality control concerning dataverses, datasets and files deposited in the research data platform of the University of Milan*
Author	*Dario Basset*

Document Status

Edition	Edition	Date
0.1	Initial Draft	15/9/2022
0.2	Added quality chapter	13/10/2022
0.3	Added requirements	18/10/2022
0.4	Added technical part	25/10/2022
0.5	Added specifications on programs	27/10/2022
1.0	Published	27/10/2022

Summary

1 Purpose

The University of Milan has created a system for storing, storing and publishing research data. The software system chosen is Dataverse (https://dataverse.org). Dataverse is a FAIR archive that provides a hierarchical structure.

Dataverse UNIMI is the platform of the University of Milan for the management of research data (https://www.unimi.it/it/ricerca/dati-e-prodotti-della-ricerca/scienza-aperta/research-data-management-lastatale).

Quality control is a complex process, requiring specific skills and tools, sometimes adapted to the context. This document, which should be understood as *a live document* (i.e. document subject to constant review), describes the guidelines for an analysis of the content of a dataverse structure, in order to ensure the quality of the metadata and the consistency of the organization and the deposited data.

1.1 Dataverse

Dataverse UNIMI is the platform of the University of Milan for the management of research data (https:// rdm.unimi.it). Dataverse (https://dataverse.org, not to be confused with Microsoft dataverse) is an open source web application to store, share, cite, explore, analyze research data, and can be used for the publication of open data.

Researchers, data producers, institutions where the same data is produced can receive appropriate credit in terms of citations thanks to the use of *Digital Object Identifier* (https://www.doi.org/) A DOI is a series of numbers, letters and symbolsused to uniquely identify a document or data, so that it is addressable with a permanent web address.

Dataverse has various installations:

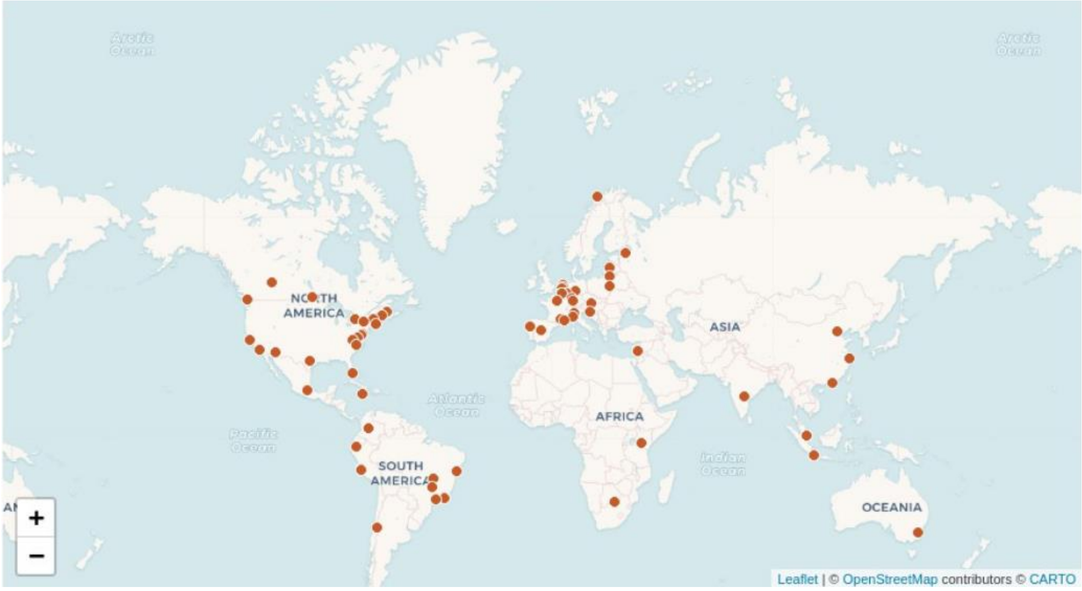

Figure 1 - Dataverse installations in the world

A Dataverse repository is an instance of the software, which then hosts multiple virtual archives called Dataverse or Dataverse collections. Each Dataverse collection can contain Datasets, i.e. file containers, or additional Dataverse collections, recursively. Dataverse, which supports the OAI-PMH protocol[1] for exchanging metadata, can make metadata of local datasets available.

All datasets created through the Dataverse interface have a default CC0 Public Domain license[2], or you can specify a custom license.

2 Introduction to open data quality

The quality of the data and metadata that describes them is of great importance for the ability to find, exchange and consult them. In the past, in the absence of a standard, digital data has suffered, and still suffers frequently, from the following situations:

- □ organized in silos (each application has its own data)
- □ different semantically (therefore not mergeable)
- □ managed by different software (therefore with different formats)
- □ incorrectly or uncertainly classified
- □ not managed with dictionaries or not yet cataloged (without metadata or without adequate semantic structure)
- □ distributed among services unevenly (without any standards)
- □ updated in an unregulated and coordinated way (therefore the actuality of the data is not known)
- □ acquired with lack of control or with low comprehensibility (the data are thus of low quality)
- □ not interchangeable, accessible and available

The normative reference for data quality is the ISO/IEC 25012:2008 standard, which has become the Italian standard UNI ISO/IEC 25012:2014, to which this document refers. To determine the quality of the data it is necessary to define measures through which to quantify the quality of the data. The standard defines a set of specific characteristics for the quality of data.

It is precisely to this standard that the National Guidelines for the enhancement of the public information heritage of the AGID refer to evaluate the quality of open public data exposed by Italian public administrations.

2.1 FAIR principles

Open research data are accessible data, even if not necessarily *open access* where there are justified needs for protection, reusable for academic, research, educational and other purposes. Ideally, open research data can be reused or redistributed without restrictions, of course it must still be verified that the license allows it and ethical, commercial and confidentiality constraints must be considered. It is logical to think that the open sharing of data increases its visibility, thus helping to build the conditions for the verification and reproducibility of research and new paths of a wider collaboration.

[1] https://www.openarchives.org/pmh/ Harvesting is the process that allows the exchange of metadata between systems.

[2] https://creativecommons.org/

The default license to use the research data, when there are no particular and justified needs for protection, must always be of free domain or at most be subject to the obligation of attribution sharing with the same open license.

Open search data must be managed according to the principles identified in the acronym FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*). The FAIR concept is quite recent, from 2014. A set of fundamental principles, called FAIR data principles, have been developed to optimize the reusability of research data. These principles represent a set of guidelines and best practices developed to ensure that data, or any digital object, is *Findable*, *Accessible*, *Interoperable*, and *Re-usable*. Below is a brief description:

☐ Findable: In order to make data reusable, it must first be traceable by humans and machines. The automatic and reliable retrieval of datasets depends on the persistent identifiers (PIDs) used, such as DOI, Handle or URN, and on the descriptive metadata attributed to the data, which must be recorded in "catalogs" or in repositories that can also be indexed by machines.

☐ Accessible: the data or at least their metadata must be accessible by humans and machines also through authentication and authorization systems (it is not necessary that the data deposited are open access) through the use of standard protocols. The data and their metadata must be deposited in archives or repositories that make them as persistent as possible over time and traceable on the network. At least metadata should always remain available even when the data is not in open access.

☐ Interoperable: Data must be able to be combined and used together with other data or tools. The data format must therefore be opened and interpretable by various tools, including other databases. The concept of interoperability also applies to metadata. For example, metadata should use a standardized language that is shared internationally by the different indexing services.

☐ Reusable: both metadata and data must be described and documented in the best possible way, to guarantee their quality and so that they can be replicated and / or combined in different contexts. The processing of data should comply with standards or protocols recognized by the relevant scientific communities. The re-use of metadata and data should be declared with one or more clear and accessible open licenses.

2.2 Curating reproducible research and FAIR: a checklist

The focus of this paper is specifically on the key issues of the care of reproducible research and FAIR data. Some issues that need to be resolved when we want to obtain reproducible research and FAIR data are:

☐ Completeness: The search compendium contains all the necessary objects to reproduce a predefined result.

☐ Organization: It is easy to understand and keep track of the various objects in the research compendium and their relationship over time.

☐ Economy: Fewer foreign objects in the compendium mean fewer objects that can break and require less maintenance over time.

☐ Transparency: The research compendium provides a comprehensive outreach of the research project that produced the scientific claim.

☐ Documentation: Information describing the objects in the compendium is provided at a level of detail sufficient to allow for independent understanding and use of the compendium.

☐ Access: It is clear who can use what, how and under what conditions, with open access when possible.

☐ Provenance: The origin of the components of the research compendium and how each has changed over time is evident.

☐ Metadata: Information about the search compendium and its components is embedded in standardized, machine-readable code.

☐ Automation: As far as possible, the computational workflow is script-based or workflow-based so that the workflow can be rerun using minimal actions.

☐ Review: A set of managed tasks necessary to ensure continuous access to and functionality of the search compendium and its components for as long as necessary.

2.3 Reproducible research and FAIR: a checklist

3 Structure of dataverses

Unimi Dataverse is designed with a basic hierarchical structure that provides a dataverse for each department, within which one (or more) dataverses can be created for each researcher.

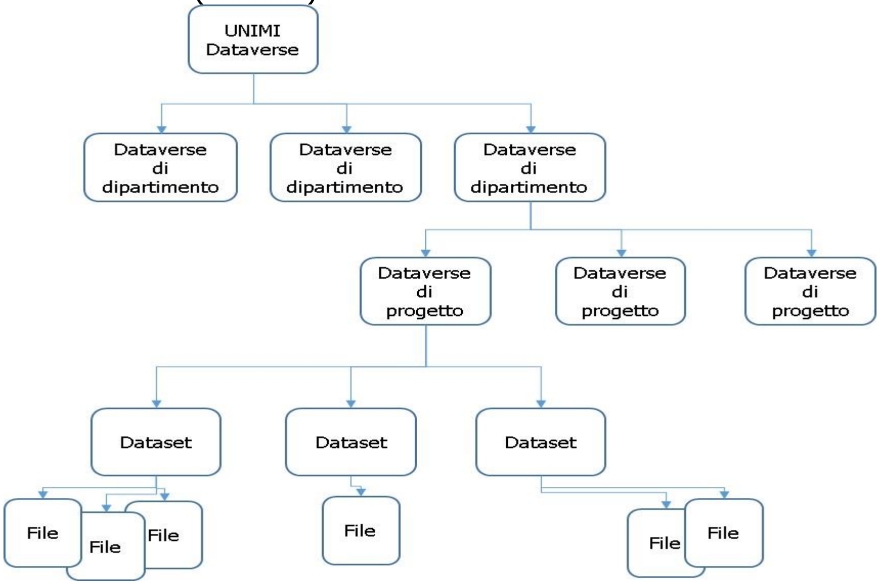

Figure 2 – structure of the dataverses

4 The main requirements analyzed on dataverses and datasets

The basic requirements to be compliant with these guidelines are of three types:

- ☐ Organizational requirements
- ☐ Metadata completeness requirements
- ☐ Completeness requirements for data access
- ☐ User requirements

Below is a brief description of each of them.

4.1 Organizational requirements

The requirements in this case are simple:

- ☐ RO01 - each user dataverse (not university or departmental) can be a project-type dataverse or a researcher-type dataverse.

Each dataverse created for a specific initiative can therefore be considered relative to a specific project or related to a specific researcher. This requirement wants to classify the user dataverses as dataverses belonging to a specific initiative, precisely a project, or, if it is not possible to trace the data deposited to a project, to a researcher. The recommended choice is normally to name a dataverse after a project.

- ☐ RO02 - each dataverse must respect the hierarchical structure of the university dataverse -> department dataverse -> project/researcher dataverses.

In practice, the requirement requires that there are no user dataverses hanging directly on the University dataverse, which sees only department dataverses at the lower level. In other words, there are no projects outside the departments. The sense of this requirement is to maintain a hierarchical structure in which each dataverse, understood as a container of other objects, respects a defined organizational structure

4.2 Metadata completeness requirements

In this paragraph we discuss the various metadata deemed necessary for a correct description of the research data deposited.

- ☐ RC01 - The basic descriptive metadata of the dataverse must be compiled entirely and accurately. Basic descriptive metadata includes:
  - ⇨ Title (**Dataverse Name**) – a unique name. It can hold spaces.
  - ⇨ Email (**Email**) – the email of the project manager or the email of the researcher must be specified.
  - ⇨ Category (**Category**) – the dataverse category must be specified
  - ⇨ Institution (**Affiliation**) – normally *University of Milan*.
  - ⇨ **Description** – the project must be described. If the dataverse contains only one dataset, simply describe the dataset.

As for datasets, there are several classes of metadata:

- ☐ *Citation Metadata*
- ☐ *Geospatial Metadata*
- ☐ *Social Science and Humanities Metadata*
- ☐ *Astronomy and Astrophysics Metadata*
- ☐ *Life Science Metadata*
- ☐ *Journal Metadata*

Of all this metadata, it is mandatory to have compiled at least the *Citation Metadata*, that is, the descriptive metadata.

Below we see in particular which fields are required filled in for *Citation Metadata* only. At least the following should be included in the *Citation Metadata*:

- ☐ RC02 - Descriptive metadata of the dataset
  - ⇨ *Title* – This is the name of the dataset.
  - ⇨ *Author* – This is a series of subfields that identify the author(s) of the project to which this dataset refers, and contains *Name* and *Affiliation*.
  - ⇨ *Contact* – The person to contact, with the Subfields *Name, Affiliation* and *Email.*

There are other optional metadata (*Alternative Title*, *Alternative URL*).

- ☐ RC03 - Dataset Subject Metadata
  - ⇨ *Description* – It's the description of the dataset, and it's critical.
  - ⇨ *Subject* – It is possible to insert a list of disciplines covered by the research (e.g. mathematics, social sciences, etc.)

Subject metadata must be compiled carefully, and there is other subject metadata, such as (*Keywords*, *Topic Classification*). In order to make the data searchable, it is suggested to also fill in the part concerning the classification. The part concerning classification, which could become important in the search/query systems that other institutions can do on the data deposited.

- ☐ RC04 - Metadata of related publications
  - ⇨ *Related Publication* – This is a non-mandatory field, but important if you know the publications related to the project. In this case, you should also enter the publication data (*ID Type, ID Number, URL*).

- ☐ RC05 - *Language* – Must be filled in. If the language of the data deposited is not English, the description must contain the reason why.

- ☐ RC06 - The metadata of those who supported
  - ⇨ *Producer* – It is a series of fields to identify the entity or administration that has the financial responsibility. The fields are Producer, Name, Affiliation, Abbreviation, URL, Logo URL, Production Date, Production Place)

- ☐ RC07 - *Contributor* – It is necessary to identify who contributed to the project. Optional.

- ☐ RC08 - *Grant Information* (*Grant Agency*, *Grant Number*) the identification of the body in charge of guarantees.

- ☐ RC09 – *Distribution data* (Distributor, Name, Affiliation, Abbreviation, URL, Logo URL, Distribution Date) the identification of the distributor.

- ☐ RC10 – *Metadata on data production* – This is a series of metadata that specify the period to which the data refers, the software used to produce them, etc.

  - ⇨ Time Period Covered
    - ○ Start
    - ○ End
  - ⇨ Date of Collection
    - ○ Start
    - ○ End
  - ⇨ Kind of Data
  - ⇨ Series
    - ○ Name
    - ○ Information
  - ⇨ Software
    - ○ Name
    - ○ Version
  - ⇨ Related Material
  - ⇨ Related Datasets
  - ⇨ Other References
  - ⇨ Data Sources
  - ⇨ Origin of Sources
  - ⇨ Characteristic of Sources Noted
  - ⇨ Documentation and Access to Sources

All this metadata on the data is not mandatory, but it is recommended in order to give more possibilities of dissemination to the dataset.

It is important to underline that at each subsequent publication the description work must be checked, as well as described the reason for the update:

- ☐ RC11 - versioning metadata must describe the rationales behind updating data

As for the files, an important trick concerns the open format. You should pay attention to the files, when the deposited files are not FAIR:

- ☐ RC12 – in the event that the data has not been produced in standard format or with non-standard tools, the data must be accompanied by a readme file.txt

Non-Fair files should be described through a readme.txt, which describes the tools used to produce them, the particular format used, and anything else that can be useful for its re-use. It is good practice to put a readme file anyway.txt, especially when the number of files exceeds 5.

4.3 Requirements for access to data (Terms Metadata)

Data access requirements require that data access policies be specified. Administrators of a dataverse can assign roles and permissions to users in the dataverse. The default access policy is the open one. In this way, unless otherwise specified, access to the data is given to everyone, of course read-only. It is also possible to share files with users not registered with dataverse, exchanging with them a link where they can see the files, for example for review.

dataverse user accounts can be assigned roles that define what actions are authorized to carry out on specific dataverses, datasets and / or files. Each role includes a set of permissions, which define the specific actions that users with that role can perform.

Roles and permissions can also be granted to groups.

Among the quality control requirements for file access, we have:

☐ RA01 – This requirement applies to datasets, for which it is possible to specify within the *terms* whether access to the specific file is free or restricted. The *terms* must be compiled in a clear and exhaustive manner, at least when there are restrictions on the files. It is understood here that the conditions of access to files, which are normally open (CC0), must be specified in case they are not. It must be written how the requested person concerned can proceed to request access to protected data, and the particular license of use if different from the standard one must also be specified.

☐ RA02 – This requirement requires that the most restrictive license be CC standard and described. The license associated by default with the datasets in dataverse and the *Creative Commons CC0 Public Domain Dedication*. If the author believes that CC0 is not suitable for their dataset, simply choose *No, do not apply CC0* and the author can enter their own terms and conditions.

4.4 User Requirements

User requirements ensure that there is at least one administrator of the dataverse and there is no proliferation of administrator roles.

☐ RU01 – This requirement checks that there is at least one administrator of the dataverse.

☐ RU02 – This requirement checks that there are no more than 5 administrators of the dataverse.

☐ RU03 – This requirement checks whether there are ORCID or internal utilities.

5 The programs

Quality control is a complex process, and the output of that process depends on the requirements initially set. In order to verify the requirements, set out in the present document, a simple programming work has been done that carries out an automatic analysis and produces some lists in output. These lists describe one by one all dataverse collections and datasets and check that the requirements expressed are met.

In order to simplify the analysis, it has been made numerical. Each dataverse collection and dataset is given a global score, an average of the individual specific scores calculated according to the requirements. The individual specific scores express the degree of correspondence of all the dataverse collections and all the datasets present, both published and unpublished. The single element (dataverse collection or dataset) that has a large point of attention is also reported.

5.1 Dataverse API

The Dataverse software suite provides a series of APIs (application programming interfaces) for the purpose of querying data on the network.

The APIs for working with dataverse software are many and offer a very high degree of completeness. It is virtually possible to replicate through APIs the entire installation of one dataverse system to another.

APIs are a tool that allows you to have detailed information on all the elements from the dataverse database, that is, on all dataverses, all datasets and all files, including those not yet published. There are also APIs that affect users. APIs can read and possibly write the dataverse database.

The tools for carrying out such queries are as follows:

- Search API with which you can make searches, even selective, or navigation (eg give me all the children of this dateverse)
- Native API with which you get detailed information about each individual element
- Data access API with which you can manage access permissions to individual files

The dataverse suite also contains other tools, for a complete discussion please refer to the official documentation.

5.2 The steps of automatic control

In the case of the University of Milan, an approach is proposed as follows:

1. Through the search API, you request the list of all dataverses, including unpublished ones. The list of dataverses contains the identifiers of the dataverses, through which we can make a call for each dataverse and obtain the detailed data of the specific dataverse.

2. The details of the fields of each dataverse are requested, scrolling through the list obtained in the previous step.

3. Through the search API, you request the list of all datasets, including unpublished ones. The list of datasets, as in the case of dataverses, contains the identifiers of the datasets. Scrolling through the list of datasets, you invoke a call with which you get the detail data of the specific dataset.

4. The details of the fields of each dataset are requested, scrolling through the list obtained in the previous step. Note that the detail data of the specific dataset also contains information regarding the files belonging to the specific dataset.

In the face of the collected data, a series of checks are carried out automatically on the dataverses, datasets and files. All the checks carried out concern the organizational, completeness and data access requirements.

The main flow of the automatic control is in charge of reviewing all the dataverses and all the datasets, in order to:

- □ Report the requirement not met in any way.
- □ Assign a score to each individual dataverse. The score will therefore depend on the characteristics of the dataverse in terms of:
  - ⇨ organizational requirements
  - ⇨ metadata completeness requirements
  - ⇨ requirements for access to data
  - ⇨ user requirements

In this way, both the anomalous situations and the degree of completeness of the dataverse are highlighted.

5.3 Automatic controls

The automatic program evaluates the **datasets by** examining the following cases:

- □ Control over the hierarchical position of the dataset
- □ Control over the completeness of dataset metadata
- □ Restricted access control
- □ Control over utilities
- □ Control over files: size and presence of the readme.txt
- □ Publisher control
- □ Control over recent activity on the dataset

To learn more

- Measuring Data Quality, AGID, 2014, https://www.agid.gov.it/sites/default/files/repository_files/documenti_indirizzo/iso_25024_agid_misurazione_della_qualita_dei_dati.pdf

- data quality: concepts and measures, Domenico Natale, Mondomatica, 2016, http://www.mondomatica.it/styled-10/downloads-3/files/NataleUNI_3mar2016.pdf

- Organizational aspects and data quality, AGID + digital team, 2014, https://docs.italia.it/AgID/documenti-in-consultazione/lg-opendata-docs/it/bozza/aspetti-organizzativi-e-qualit%C3%A0-dei-dati/qualit%C3%A0-dei-dati.html

- Systems and software quality requirements and evaluation (SQuaRE) – Measurement of quality in use (25022:2016), ISO, 2016, https://www.iso.org/standard/35746.html