UNIVERSITÀ
DEGLI STUDI
DI MILANO

LA STATALE

Institutional metadata & data quality check: experiences

GenOA week 2022

Direzione Performance, Assicurazione Qualità, Valutazione e Politiche di Open Science
University of Milan

Quality of data experiences

Contents:

- ❑The problem of research data & metadata quality
- ❑What is quality
- ❑Dataverse/Dataset structure
- ❑The semi-automatic check

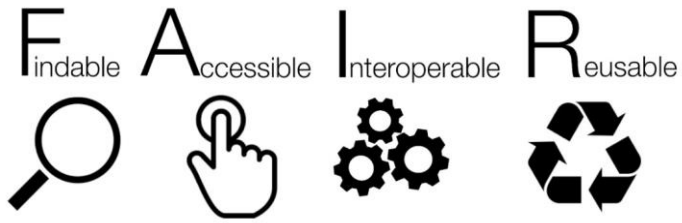
The problem

Problem statement

Data quality is critical with respect to FAIR concept.

Metadata quality is critical to make research data FAIR

The University of Milan has chosen a data repository fully compliant with FAIR



The problem

The problem

Researchers are the depositors of data and metadata in [Dataverse UNIMI]

How repository staff can check repository FAIRness?

Dataverse
UNIVERSITÀ DEGLI STUDI DI MILANO

Situation

Depositors + **Datasets** + **Staff** =

Actions

Users

- [context]
- [guidelines]
- [training]

Repository

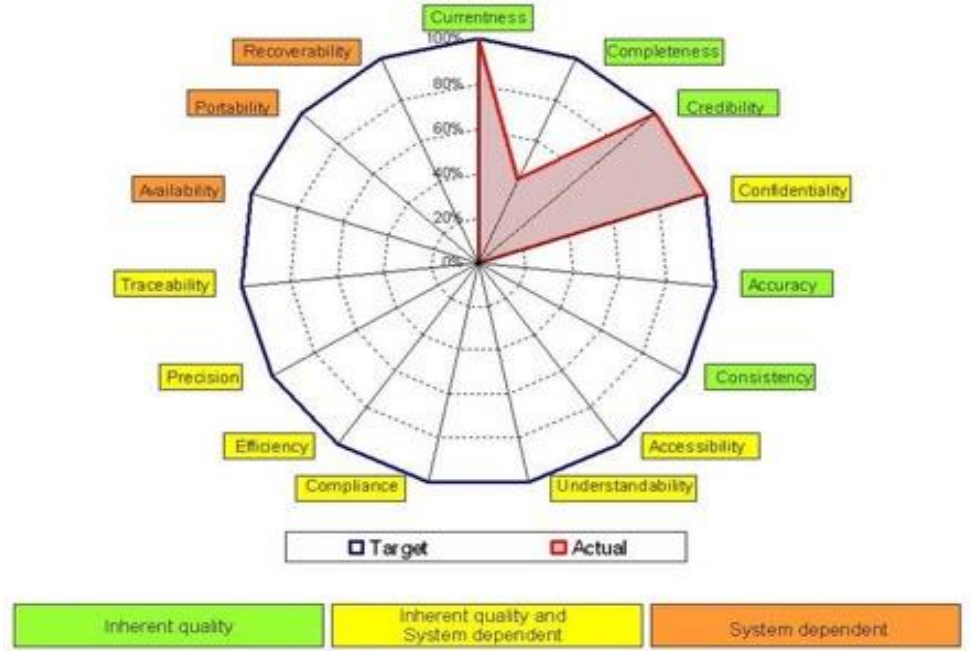
We are working on semi-automatic control of published data and metadata

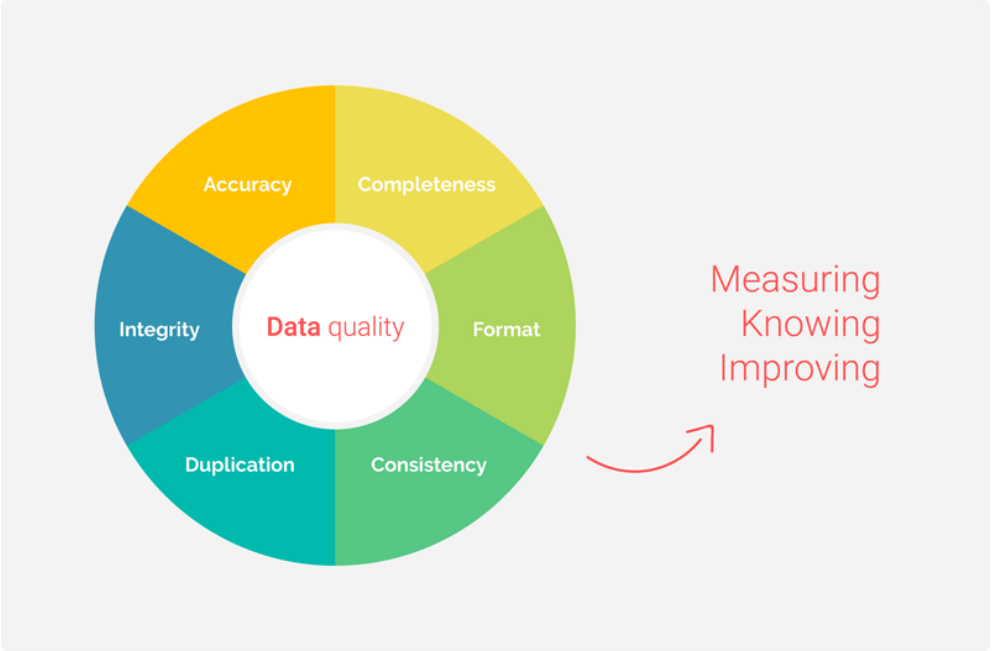
Quality definition

Data Quality Definition

Data quality is the measure of how well suited a data set is to serve its specific purpose. Measures of data quality are based on data quality characteristics such as accuracy, completeness, consistency, validity, uniqueness, and timeliness.

As expected, data quality depends on its use!





Quality definition

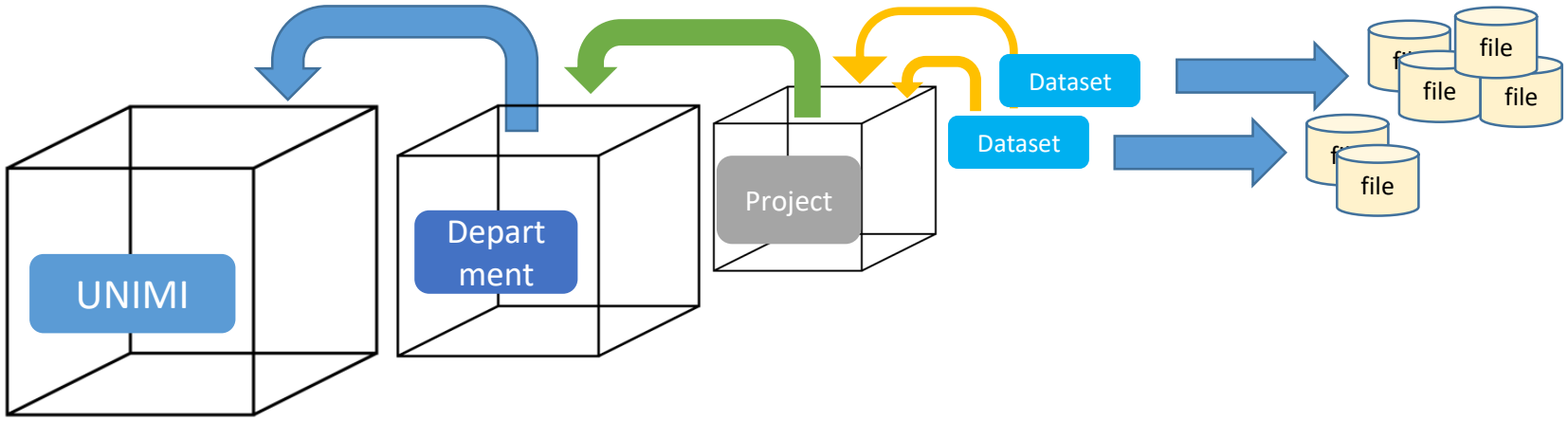
Data Quality ISO/IEC 25012

There are six main dimensions of data quality: accuracy, completeness, consistency, validity, uniqueness, and timeliness.

- **Accuracy**: The data should reflect actual, real-world scenarios; the measure of accuracy can be confirmed with a verifiable source.
- **Completeness**: Completeness is a measure of the data's ability to effectively deliver all the required values that are available.
- **Consistency**: Data consistency refers to the uniformity of data as it moves across networks and applications. The same data values stored in difference locations should not conflict with one another.
- **Validity**: Data should be collected according to defined business rules and parameters, and should conform to the right format and fall within the right range.
- **Uniqueness**: Uniqueness ensures there are no duplications or overlapping of values across all data sets. Data cleansing and deduplication can help remedy a low uniqueness score.
- **Timeliness**: Timely data is data that is available when it is required. Data may be updated in real time to ensure that it is readily available and accessible.

Dataverse

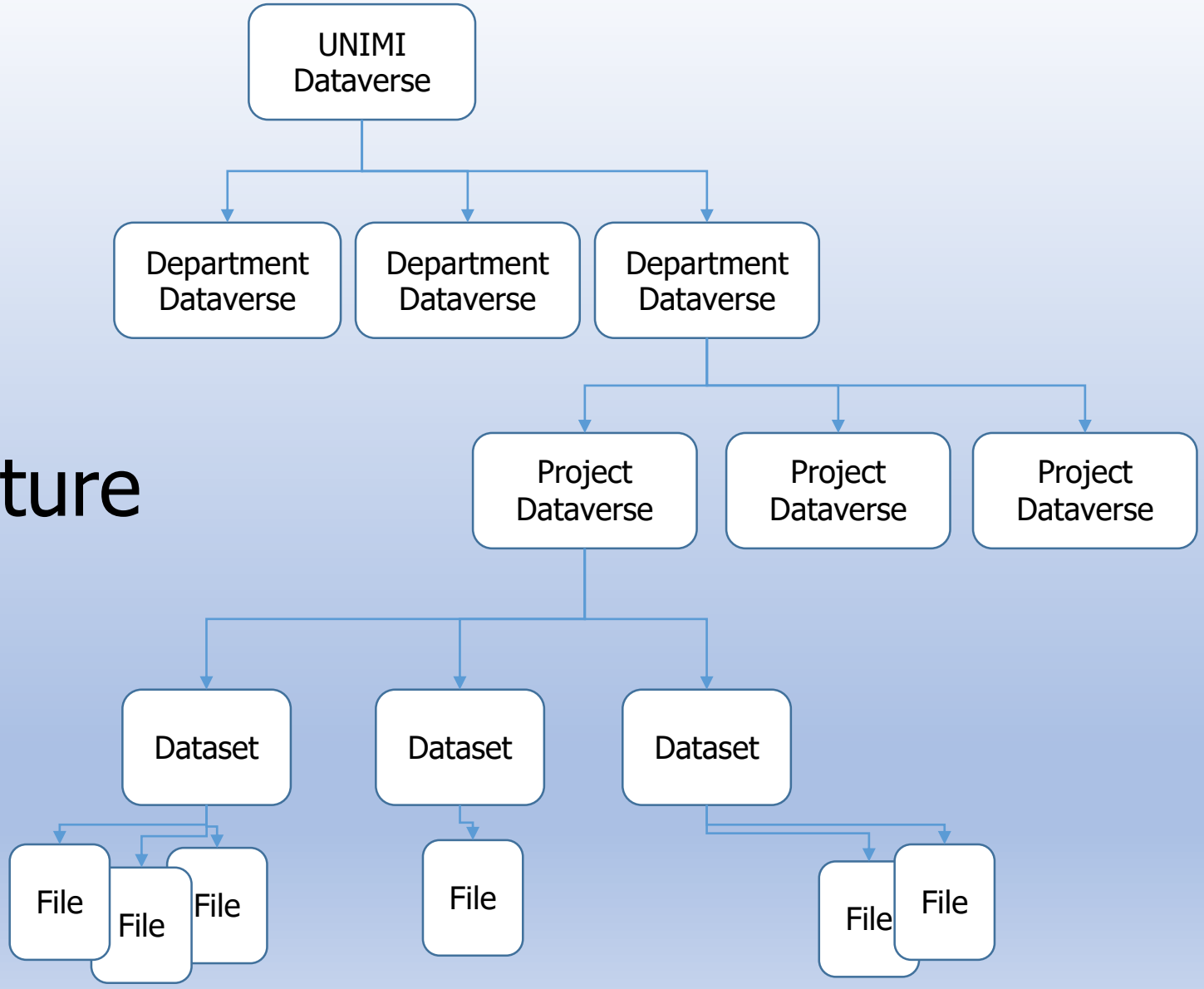
Dataverse UNIMI *(https://dataverse.unimi.it/)* is the institutional platform for sharing and managing research FAIR data



- ❑ Dataverse UNIMI structure implies a dataverse for each department.
- ❑ Inside the department dataverse, each project or researcher can have Dataverse.
- ❑ Lastly, inside the project/researcher dataverse there are the Datasets.

Dataverse

A tree structure



The check

The steps:

- ❑Define which are the main requirements
- ❑Set a priority for each requirement
- ❑Run the automation
- ❑Produce the list of the datasets subject to review

The check

Main dataset requirements

- ☐ **Organization requirements** – check for correct positioning of the dataset inside the tree
- ☐ **Completeness requirements** – check for metadata completeness. Check for README existence
- ☐ **Data access requirements** – check if the access to data is restricted and check if correctly specified
- ☐ **User requirements** – check if there are admin users and if the dataset is correctly administered

The check

The program – tech specifications

- ❑ **Batch PHP program.** It is conceived as a batch program that can be launched on need. Since it reads Dataverse API, the program can be launched from any server or even from a PC.
- ❑ **Will be deposited in GITHUB in Q1 2023.** Repository is https://github.com/joelfan/Dataverse-quality-check.
- ❑ **Customizable.** There is a config file that is to be customized with all context parameters and execution parameters.
- ❑ **The PHP uses dataverse API to get information.** Dataverse offers a complete set of API that can be queried to obtain all relevant data.
- ❑ **Only published datasets at the moment.** The analysis is conducted through datasets and related files.

The check

The results

- Results are produced in txt, csv and xlsx files. A list of published datasets is produced with some attributes for each dataset.

DATASET Id=0000 Name=https://doi.org/10.13130/RD_UNIMI/xxxxx
Publisher=UNIMI Dataverse
Creation date=2020-08-13T06:08:23Z
Version status=DRAFT
Publication status=Published
Publication date=2019-01-16
License=CC0
Terms of use=CC0 Waiver
DOI=https://doi.org/10.13130/RD_UNIMI/xxxxx
dataset Persistent ID=doi:10.13130/RD_UNIMI/xxxxx
# of files=2
Total size of files in MB=0,008
Readme=YES
Entity ID= 0000
Global ID=doi:10.13130/RD_UNIMI/xxxxx
Identifier of container Dataverse= fatherDataverse
Name of container Dataverse= fatherDataverse
Dataset contact=OK
Restricted=NO
File access=
Last update=2020-08-13T06:08:23Z
Creation=2020-08-13T06:08:23Z
MSG=
ATTENTION=

*** Score ***	Pos	Cit	Siz	Act	Pub	Acc
	4	4	4	4	4	4

Organization – correct positioning of the dataset inside the tree

Completeness - metadata completeness. README

Data - readme & data size

Data - update activity on data

Completeness - publication

Data access - any restrictions

UNIVERSITÀ DEGLI STUDI DI MILANO
LA STATALE

The check

The results

- Published datasets: **162** (out of 288)
- Number of datasets with severe issues: **0**
- Number of datasets with warnings: **31**

Most of warnings are related to readme existence and unspecified terms of access

UNIVERSITÀ DEGLI STUDI DI MILANO
LA STATALE

Q&A