

Session 3: Systems and processes underpinning Real-World Data

Characterisation and maturity model consideration

Presented by Ana Cochino (EMA)

Multi-stakeholder workshop on RWD quality and experience in use of RWE for regulatory decision-making – 26 06 2023

An agency of the European Union



In the EU DQF, we partition data quality aspects ("determinants") in three different categories

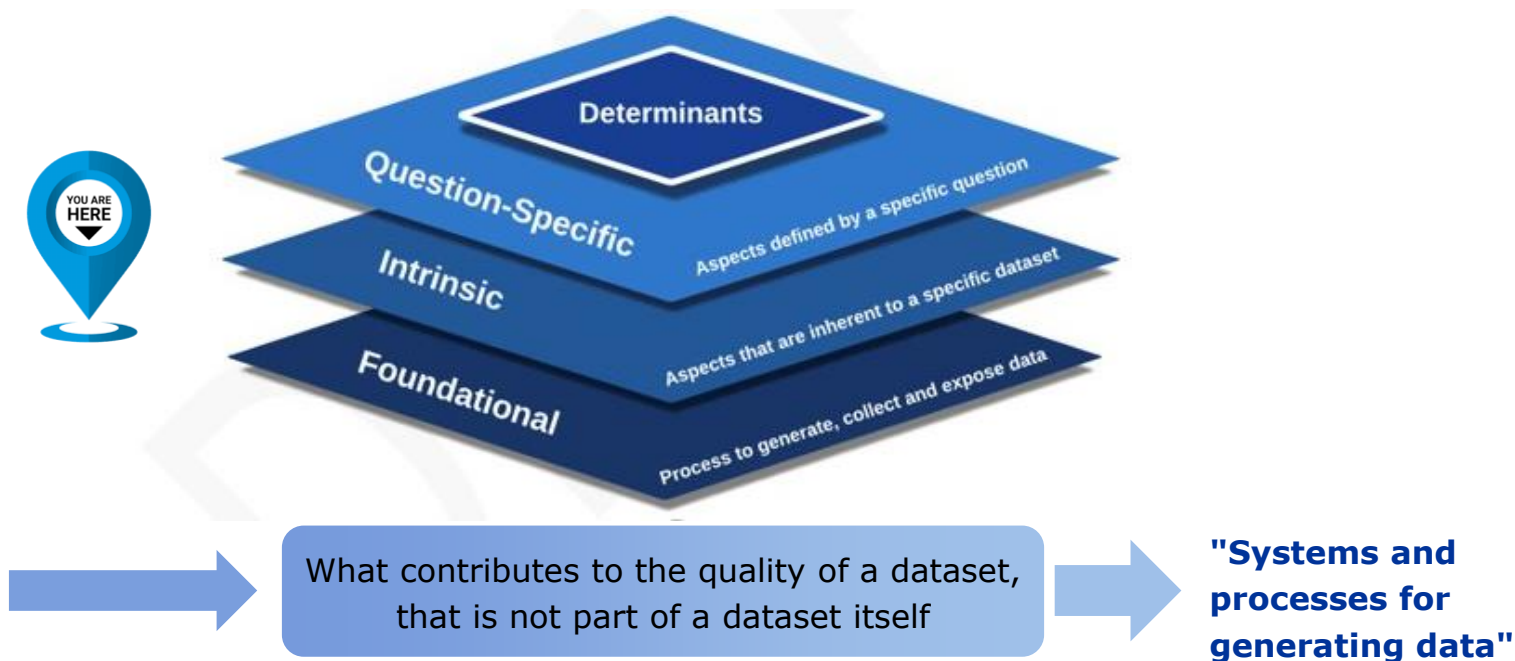


What pertains to quality that cannot be assessed or measured without a defined research question and method

What quality aspects can be assessed or stated about a dataset, independently from the way this was generated

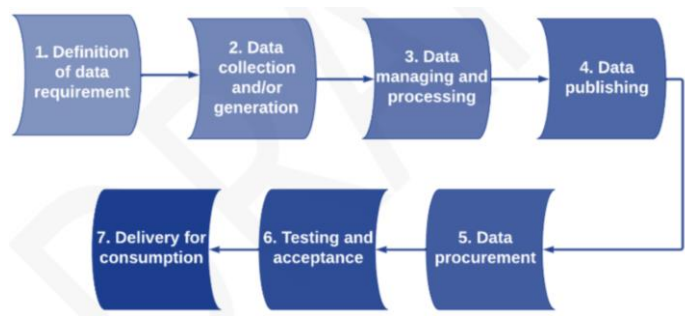
What contributes to the quality of a dataset, that is not part of a dataset itself

In this session, we focus on foundational determinants of quality



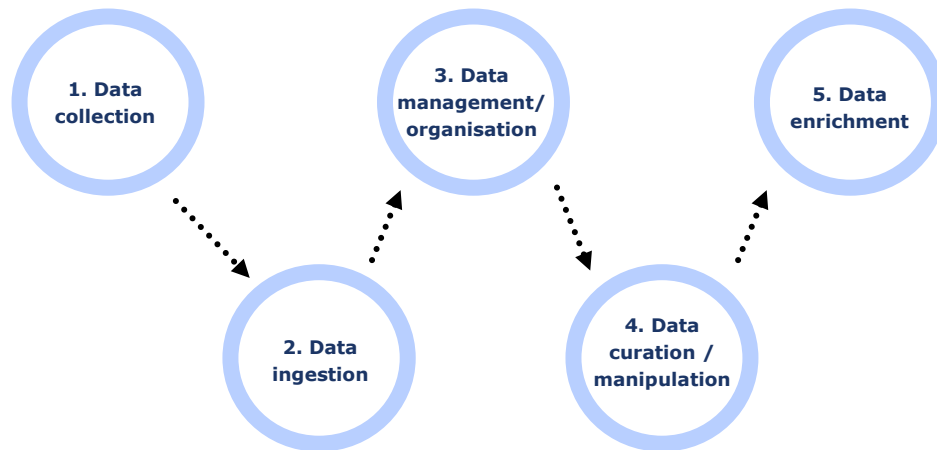
RWD is characterised by different data generation workflows, where each step has an implication on data quality

Typical steps in general data flow (as seen in general DQF)



Example steps in RWD flow with quality implications

based on RWD specificities (e.g. possible fragmented collection and ingestion, data transformation to standards, linkage, etc)



Maturity model for the characterisation of systems and processes

The maturity model levels below propose an overall categorisation of the systems and processes generating data

1 Documented

- A basic level of documentation of relevant steps (SOPs, data transformations,...)

2 Formalised/standardised

- The documentation (and the related DQ practices) are not created ad hoc, but based on shared standards and processes, to help its interpretability


3 Implemented

- Beyond using shared standards and approaches, the DQ practices are implemented in processes so that they are collected and correct by design

4 Automated

- Quality assessment is machine readable, as to be able to enable automated feedback loops to improve data quality

The systems and process characterisation checklist

Process step or data source element	Goal of characterisation	Characterisation to be provided
Rationale and scope for data collection	Provides an overall understanding of data quality strengths and biases. Provides benchmarks and comparators	
Original source of data	Provides understanding on the origin of data, including impact on reliability and accessibility of various factors (e.g. limited human or material resources or inadequate training) influencing quality	
Data collection/ ingestion methods	Demonstrates robustness of the data collection process	
Data management practices and infrastructure	Informs of data management practices ensuring reliability of data	
Data manipulation/ curation steps	Describes reliability of data in relation to manipulation and transformation steps, verification and monitoring of data cleaning, extraction and transformation processes	
Data enrichment steps	Describes reliability of data in relation to enrichment (manipulations that include the alteration of data, such as imputations of missing values)	
Known quality issues	Demonstrates awareness of systematic quality issues and increases transparency	
Data source description*	Provides the data dictionary and standards used to the "published" version of a dataset (* not strictly part of a systems and processes characterization, but pragmatically useful to present in the context of a checklist)	

RW data sources: EMA/HMA Catalogue of data sources (metadata catalogue)

- The '**metadata list**' describing RW data sources has been developed with the primary objective of promoting 'data discoverability';
- The data elements (metadata) proposed are used in the **implementation** of a publicly available **catalogue of real-world data sources** meant to:
 - Help regulators, researchers and pharmaceutical companies to **identify data sources suitable** to address research questions, based on the 'FAIR' (findable, accessible, interoperable and reusable) data principles
 - **Boost transparency of observational studies** (via the linkage to the catalogue of studies)
 - Improve the ability of the relevant stakeholders to **assess evidence** from observational studies and real-world data sources

RW data sources: EMA/HMA Catalogue of data sources

It captures (broadly) the following characteristics of a data source:

- Administrative details: name, data holder, geographical coverage, type of data source (..)
- Data collected: specific diseases, sociodemographic information, lifestyle factors (..)
- Data governance : overall governance of the data source and processes and procedures for data capture and management, data access, data quality check and validation results
- Possibility to audit data and external validation
- Data lifecycle: events triggering registration/de-registration
- Linkage behind the creation of a data source (where applicable)
- CDM mapping, ETL status
- Capturing of structured data using standard terminologies

Process step or data source element

Rationale and scope for data collection

Original source of data

Data collection/ingestion methods

Data management practices and infrastructure

Data manipulation/curation steps

Data enrichment steps

Known quality issues

Data source description

Points to consider

- The checklist is intended to provide a **characterisation of the different steps** in the data generation process, so that the impact on data quality can be better understood and assessed.
- Such characterisation can be described in different forms:
 - **Checkbox:** a system/process/procedure exists
 - **Submission of relevant documentation** or web reference
 - More structured and **formalised information** (e.g.: split of information in structured fields).
 - **Automated integration** of metadata elements and update
- The defining of a "maturity model" is meant to **describe the Data Quality related systems** and processes 'at a glance' and to encourage towards automation and interoperability
- The metadata elements proposed for Data Source Catalogue are a good base for building the **process and system characterisation** (= data elements that were agreed as being useful in characterising a data source)

Proposed discussion points

- Implementation of a (more) **extensive data quality module** as part of Data Source Catalogue – pros and cons?
- The Data Source Catalogue contains publicly available information – would this pose any **challenges in collecting proposed data** quality information?
- (For the RWD user) What is the **most important information** to have in terms of processes and systems when assessing the quality of a data source?
- (For RW data source holder) **How effort-intensive** would you characterise the submitting of the information proposed?