Data (dataset) Quality Control Questions

To ensure the quality of a research dataset, it is important to evaluate the data against a variety of criteria.

The control questions in this list are based on the most important principles, organized into logical groups that cover data accuracy, availability, security, compliance with the FAIR principles, and open data requirements.

The questions are primarily intended for reviewing datasets in cases where it is decided to deposit or store them in a data repository.

1. Accuracy and completeness of data

Accuracy

- Does the data correspond to reality and are free of errors?
- Has the data been checked for errors (e.g. duplicates, incorrect entries)?
- Are the data values consistent across all records?

Completeness

- Is the data complete and covers all the information necessary for the study in a particular case?
- Is the impact of missing data on the study documented and understood?
- Does the dataset contain all the files?

2. Availability and openness of data

Access

- Is the data accessible through standardized public access protocols (e.g. HTTP, FTP)?
- Are there clearly defined access conditions (free access, restricted access)?
- Is the data the best possible solution for preservation and/or long-term preservation?

Long-term availability

- Is the data available in the long term through trusted data repositories or archives?
- Have steps been taken to ensure that the data will continue to be usable in the future?

Findability

- Is the data described with metadata and searchable in a repository or other information system/s?
- Is the data described using the metadata standard of the relevant field of science?
- Are metadata records in open access?
- Does the metadata contain unique identifiers?
- Do metadata entries use the relevant discipline-controlled dictionaries, including ontologies, to create keywords?
- Is the organization of datasets using high-quality file structuring and versioning?

Openness

- Is the data available without legal restrictions?
- Is the data available without financial restrictions?
- Is the data available without technical restrictions?
- Is the data available in open, standard formats (e.g. CSV, JSON)? If not, then what's the rationale?

3. Data documentation and transparency

Interpretability

- Is the data sufficiently documented to be understood and used by other researchers?
- Is there a ReadMe file attached to the dataset?
- Is the data described with sufficient metadata (data structure, meanings, encodings)?
- Does the code book/dictionary reflect all the variables used in the data and their values?

Transparency

- Is the data collection and processing process clearly documented and easily traceable?
- Is documentation available on data processing and methods used?

Traceability and origin

- Is full documentation on the origin of the data and the history of changes attached?
- Have any changes made since the start of data collection been documented?

Tools used

- Is the version and configuration of the software or tools documented?

4. Data compatibility and reusability

Interoperability

- Is the data compatible with other datasets using standardised file formats and protocols?
- Does the data use internationally recognised standards (e.g. ontologies, classification systems, methodologies)?

Sharing rights

- Who owns the data? Who has the right to decide to share them?
- Are the data consistent with the policies of the institution representing the author of the dataset?
- Has there been a balancing of the institutions' policies/agreements in the case of a multi-partner research project?
- Will the data be in line with the research funder's policies for the dissemination and sharing of research results?

Reusability

- Is the data documented and licensed in such a way that it can be reused in different contexts?
- Are there conditions for reuse (e.g. what type of Creative Commons license)?
- Does it specify how long sharing will be provided (including from what point in time, or will the embargo period be used)?
- Has a description of potential data usage needs and/or users been added?
- Is it specified how the dataset should be referred to in the case of use?

5. Data security, privacy and ethics

Privacy protection

- Whether the data complies with international data protection standards (e.g. General Data Protection Regulation (GDPR)?
- Has the data been anonymised if the data contains personal information?
- Has the data been pseudonymized if the data contains personal information?

Data security

- Are security measures in place to prevent unauthorized access and protect data from theft or damage?

Ethical requirements

- Is the data managed in accordance with ethical guidelines, especially for sensitive information?
- Is appropriate consent from study participants, if necessary, obtained?

6. Reproducibility and reliability of data

Reproducibility

- Could other researchers reproduce the results of the study based on the available data and documentation?
- Are the methods of data collection, cleaning and analysis fully documented?
- Are the versions of the dataset clearly indicated and the changes between versions documented?

Reliability

- Has the data come from reliable sources and has its reliability been verified?
- Are the circumstances of the time and place of data collection indicated to ensure their suitability for the study?
- Are the tools and tools used adequate, reliable/validated and documented?
- Are the versions and configurations of the software or tools used documented?