



CHECKLIST OF QUESTIONS FOR RESEARCHERS WORKING WITH DATA

This checklist aims to help researchers navigate through the maze of challenges, opportunities, restrictions and obstacles related to storage and availability of data for research, and to help them secure the necessary support from the home knowledge institution. Data are taken to be digital data.

By following the checklist below, the researcher becomes aware of the importance of proper storage of and access to data, and of the steps to be made along the path of data generation and analysis. In some of these steps, the researcher needs support from his/her employer. Making clear what support is needed and how it is realised can be a major challenge. It must become clear in advance of the research project who is responsible for providing and organising proper data storage and access.

Frame

Data come in different categories:

- Data that are indispensable for other researchers.
- Data that are useful for other researchers, but not indispensable.
- Data that are potentially useful in the future.
- Data that do not need to be stored and for which it suffices to describe how the data were generated.
- Data for which reproducibility or complementary studies are the main reason for storage.

Important aspects of storage and availability of data for research are:

- Data are useless without metadata. Metadata are necessary to make data findable and (re)usable for other researchers. Metadata contain information about how the data are generated, organised, processed and stored, including software and analysis scripts.
- Data may be eliminated in the course of the research process, or thereafter. In case of personal data, this may happen at the request of the data donor and on the basis of his/her legal right to be forgotten.

Questions

The following questions need to be addressed before starting the research project:

(A) Which data are allowed to be generated, which data are allowed to be accessed, and by whom?

Sub-questions are:

- 1 Both the generation of data and the access to data may be restricted. What restrictions apply at the home institution and nationally (and possibly internationally)?
- 2 How do users authenticate themselves before they are granted access?

Amsterdam, May 2021



- 3 Are there any legal or ethical hurdles to be taken?
- 4 What agreements need to be made beforehand and with whom? Are there contradictory requirements to be bridged?

(B) Which data and support will be needed for a proper execution of the research?

Sub-questions are:

- 1 Of the necessary data, which will be primary data collection and which secondary data collection? Data can be collected either by the researcher (primary data collection) or by others (secondary data collection). In case of personal data, the collection of primary data should comply with the General Data Protection Regulation (GDPR). Does the collection of secondary data comply with the conditions of the data source, including possible financial compensation and possible sharing of responsibility for the data?
- 2 Are the data made FAIR? How is this measured? What are the standards?
- 3 Does the home knowledge institution offer enough facilities and support to acquire the necessary data? If no, then where can this be secured? If yes, then what are the local rules for data storage, the local support points and the local support tools? What are the protocols? Who is responsible for what? Are training and workshops being offered? How do local initiatives fit with national plans? Who are the local legal experts, ethical experts, data stewards and software engineers?
- 4 Which preparations need to be made prior to the collection of primary data? What needs to be put in a consent form?
- 5 Who needs to be contacted prior to the collection of secondary data?

(C) Which data will be stored and kept for future access, and which will not?

Sub-questions are:

- 1 If data are stored and kept for future access, then why is this so? What are the costs and the benefits involved? What if the data volume is so large that storage costs are prohibitive? How long will the data be stored and why? Who is the owner of the stored data? What are the roles of the researcher, the home knowledge institution, the granting agency and, in case of personal data, the data donors? Is there sufficient agreement on ownership and roles? What are the standards for storing and keeping data for future access? Is there coordination?
Is there transparency? Will the data be stored and kept for future access with a Digital Object Identifier (DOI) and with metadata? Is there an Open Science community with whom experiences can be shared and discussed? Are there national or international data infrastructures to make use of? Are there colleagues with relevant expertise who can help, for example legal experts, ethical experts, data stewards and software engineers?
- 2 If the data are not stored and kept for future access, then why is this so? In case of personal data, are there privacy issues? Is there a lack of storage space? Are the costs involved too high? Are the data difficult to classify?

(D) Will data be made accessible for reuse by other researchers?

Sub-questions:

If the answer is 'no', then:

- 1 What are the motives for not making data accessible for reuse by other researchers? For instance, can data not be made accessible because of privacy issues, ethical matters or restrictive contracts? Can these obstacles be removed? What to do when circumstances change along the way?
- 2 Will data be made accessible later? Perhaps data can be made accessible after some pre-specified delay or embargo.

If the answer is 'yes', then:

- 1 What are the motives for making data accessible? Possible motives are:
 - Funding agencies (such as the home knowledge institution, NWO, ERC) require that data are stored.
 - There are incentives for the researcher to make data accessible, such as long-term financial support.



- Other researchers are more willing to collaborate after data have been shared. Collaboration can improve data.
 - Other researchers are more willing to quote research publications or refer to data that are made accessible.
 - Data are indispensable for other researchers.
 - Data have potential value in terms of reuse (quality, originality, innovation, size, scale, cost), uniqueness (non-repeatable observations), or history (in particular, history of science).
 - While not indispensable for other researchers, data are indispensable for complementary studies.
 - Data are essential or useful for non-scientific purposes, for example, cultural heritage or museums.
 - The Netherlands code of conduct for research integrity (VSNU et al., 2018) states that data must be stored and made accessible for reuse, so that in the future research can be verified and reproduced. From the viewpoint of science integrity, others should be able to verify the research findings, and to check the data to avoid fraud. Not only raw and published data must be stored properly, but processed data as well.
 - Academic publishers require that the published data are made accessible for reproducibility studies. They provide researchers with facilities and support for storage and accessibility of the data.
- 2 Which data will be made accessible? Possible forms are:
- Make all or part of the data accessible. Processing of personal data must comply with the GDPR.
 - Data include suitable metadata necessary for analysis.
 - Software and analysis scripts are also data.
 - In case of controlled access to data, the researcher adds metadata to search catalogues.
- 3 To whom will the data be made accessible? Possible target groups are:
- All researchers.
 - Specific groups of researchers.
- 4 At what moment will the data be made accessible? Possible moments are:
- Immediately upon completion of the research project.
 - At some later time that has been agreed on beforehand.
- 5 In what form and under what conditions will the data be made accessible? Possible forms are:
- Decide on file formats and code systems to represent the data in such a way that other researchers can easily understand and process the data.
 - Decide on the conditions for making data accessible, including financial compensation for relevant costs. These can include the costs of the provision of data, data storage, and data transferral. The potential re-user shares responsibility for the data.
- 6 What is the best form to transfer the data? Possible forms are:
- Transfer data either directly, via a data journal, or via a local, national or international repository or archive.
 - In case of personal data, data analysis (for example, meta-analysis) may be carried out within the walls of the home institution of the owner of the data (so that the data do not need to go out), either automatically or manually.

This checklist is part of Annex 2 of the advisory
report: KNAW (2021) *Storage and availability of data
for research. From intentions to implementation.*
Amsterdam, KNAW