Amsterdam University of Applied Sciences

Data package checklist Urban Vitality

Niek van Ulzen, Amsterdam University of Applied Sciences
Version 1, 01-07-2020

The ultimate goal of FAIR is to optimize the Reuse of data. To achieve this, metadata and data should be well-described and documented so that they can be replicated, understood and/or combined in different settings. Think of variable labels, codebooks, protocols and instruments used, attaching a license, etc. This checklist details what to include in a data package besides the data itself. The data package can be deposited in data repository such as UvA/HvA figshare.

Notes and prerequisites

- The data package is about archiving research data to facilitate data reuse and verification and replication of results
- The data package is about (static) research data underlying a scientific publication or report
- The data package is not about archiving all 'essential project documentation'
- In terms of FAIR the data package is mostly about increasing the human readability of the data and to a lesser extent about machine-readability
- The data package is provided in English
- The data package should be deposited in a data repository *before* publication of results in a scientific journal. In the article the *doi* of the data package (provided by the data repository) can be referenced
- Try to use open, preferred formats for the data package as much as possible. See here a list of preferred formats specified by DANS
- (Raw) datasets don't contain directly identifying information (such as name, (e-mail) address, postal code, etc.). This information is stored in a linking table or key file (a file containing directly identifying information that is linked to research data via a random id code) and should be archived (if necessary to archive) separately from the data package. Ask opensciencesupport@hva.nl for more information

This guideline is based on:

1) HvA RDM richtlijnen 2018 (intranet, Dutch), article 8-12
2) Netherlands Code of Conduct for Research Integrity 2018, article 3.3.25, 3.4.25, 3.4.45 and 4.4
3) The data package checklist is largely based on Guidelines for the archiving of academic research for faculties of behavioural and social sciences in the Netherlands and the local implementation guidelines of the Faculty of Behavioural and Movement Sciences
4) GO FAIR website, FAIR principles. https://www.go-fair.org/fair-principles/
5) Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Acknowledgements
This checklist is one of the results of the Urban Vitality project 'FAIR: geen woorden maar data' made possible by a Urban Vitality seed money grant and by a SIA SPRONG grant for 'Mensen in Beweging'.

Archiving component	Remarks
A brief description of the problem definition, research design, conceptual framework, data collection (sampling, selection and representativeness of informants) and methods used	- Not required as long as the manuscript or study protocol provide this information in detail - If previously archived, documented in a published preregistration or described in detail in a published manuscript, cite the persistent identifier (location) of this information
The instructions, procedures, the design of the experiment and materials that can reasonably be deemed necessary in order to replicate the research	- For example, topic list, interview guide, questionnaires, standard operating procedures - The materials must be available in the language in which the research was conducted - If previously archived, documented in a published preregistration or described in detail in a published manuscript, cite the persistent identifier (location) of this information
Raw (pseudonymized) data files	- For example, a locked Castor database, export file of an online survey, recordings or transcripts of interviews, descriptions of observations, a snapshot of databases that constantly change (e.g. registries or long-term cohorts) - If the raw data is already archived elsewhere, provide a reference to where the archived raw data can be found - The raw data should not contain directly identifying personal information - The variables and values of the raw data file should be labeled and/or described in a codebook/data dictionary
Analyzed data files	- The processed data files that were eventually analyzed when preparing the article (e.g. an SPSS data file after transforming variables, after applying selections, etc.) - Not required if the raw data file was directly analyzed and the computer code is provided (see below) - It is important that the correct version of the file is submitted - Any intermediate files created during the process of raw data into analyzed data do not need to be archived, as long as the code showing the processing steps are provided (see below)
Computer code and syntax	- For example, Atlas.ti, SPSS syntax file, R code, Python analysis script - Computer code raw data file > analyzed data file. Describing the steps taken to process the raw data into analysis data, including brief explanations of the steps in English, for example a brief description of the steps taken in the qualitative analysis of research data, i.e. themes, domains, taxonomies, components - Computer code of the (statistical analyses) - Computer code analyzed data file > manuscript results (e.g., figures, tables, etc.). Describing the steps taken to process the analysis data into results in the manuscript, including brief explanations of the steps in English - Version control parameters should be provided and the correct version of the code must be submitted with the archiving package - The computer code should 'run' (work) on the provided raw and analyzed data files
Additional documentation and metadata necessary to replicate the research or reuse the data:	
Codebook / data dictionary	- A data dictionary is critical to making your research more reproducible because it allows others to understand your data. The purpose of a data dictionary is to explain what all the variable names and values in your spreadsheet really mean. See here and here. - For example, units used, definition and labels of (categorical) variables, codes for missing values, description of how derived variables were created, allowed values
A readme file describing which documents and files can be found where and how they should be interpreted	The readme file must contain (part of) the following information: - Name of the person who stored the documents and files - Date, period, location of data collection - Names of people who collected data - How the various files are related to each other and where related documents, data, preregistrations, protocols can be found - Descriptions of measurement instruments used (device, brand, version, calibration procedures) - Hardware and software used (including version) - Description of quality assurance (checking for errors, validation methods) The readme file should be saved in a non-proprietary format, such as .txt or .xml
Statistical Analysis Plan and Data Management Plan	If previously archived, documented in a published preregistration or described in detail in a published manuscript, cite the persistent identifier (location) of this information
Documents relating to the ethical approval	- The empty informed consent form (information sheet + consent form) should be archived because it shows what participants consented for (e.g. data reuse or not) - The signed informed consent form should be archived separately from the archiving package. In case of paper informed consent, see non-digital archiving
Logbooks or lab journals	Notes taken during the project