Explanation of the readme template for datasets

Hilda van der Pol
Version 2023-03-06

The readme file is a document that accompanies the dataset. It contains information (metadata) on the dataset. The metadata is also added to the description of the dataset in the archive. The readme file also provides an explanation of the dataset, making the dataset understandable and reusable.
This document explains the different sections of the readme template provided by Wageningen University & Research - Library. Some sections may not be applicable to your dataset; you can delete these sections from the file. Save the template as README.txt.

Title of the dataset:
Provide a title that reflects the content of the dataset. If the title of the dataset is the same as the title of your publication, you can add 'Data from:' or 'Data underlying the publication:' to make the dataset easily distinguishable from the publication.

Creators:
People who created the dataset or made a major contribution to it. Creators are part of the citation of the dataset. Include the affiliations and preferably also the ORCIDs.

Contributors:
People who made a minor contribution to the dataset, i.e. a laboratory assistant, supervisor or project leader. Contributors are not part of the citation of the dataset. Include the affiliations and preferably also the ORCIDs.

Related publication(s):
The paper(s) based on these data. Usually one, but it is possible that several publications are based on the same dataset. A dataset is a publication on its own and can be published without a related publication. If your paper is not accepted yet, this information can be added later to the description in the archive.

Description:
A short description (abstract) of the research (2-3 sentences). This is also part of the description of the dataset in the archive.

Keywords:
Keywords are added to the description in the archive. Keywords improve the findability of your dataset. Use both general and specific keywords.

Spatial coverage:
The country or region where the data was collected, i.e. Kenya or Gelderland. If this is not applicable to your data, i.e. because you performed an experiment in a laboratory, you can delete this section.

Temporal coverage:
The period during which the data was collected (e.g. 2007-2010) or the period the data applies to (i.e. the Middle Ages). If this is not relevant for your research, you can delete this section.

Files:
List the files names. If necessary, you can add an explanation of each file.
If you have a large amount of files, state the number of files and give an explanation of the file names.
If your files are organized in folders, explain what each folder contains.
The purpose of this section is to provide a future user of the dataset enough information so he/she can determine whether the dataset is complete and he/she can understand what the files contain.

Explanation of the variables:

If your dataset contains tabular data, explain the variables in this section, including the units of measurement.

If you have a large amount of variables, you can put this information in a separate file. Use a sustainable format, i.e. txt or csv. Add this file to the list of file names and mention here that there is a separate file with the explanation of the variables.

Methods, materials and software:

This section should include all information on how the data was collected or generated, thus making it possible for a future user to determine if the dataset can be used for his/her purpose.

The content of this section can be very diverse. Each research project and each dataset is different. You can think of:

- The strain or species used in the experiment
- Protocol of the data collection
- Brand and type of the measuring instrument(s)
- Type and version of the software
- Source code and where it can be found if it is not archived with the dataset

If this is a lot of information and it is already in your publication, you can refer to the publication. Please note that if your publication is not Open Access, people might not be able to get the information.

Licence:

For Open Access publishing of your dataset, the Library advises the CC BY licence, but other licences are possible. For more information on CC-licences, see https://creativecommons.org/about/cclicenses/

Note: Simple representation of factual data cannot be protected by copyright and the CC BY licence does not apply to that part of the data set. However, attribution is standard good research practice and is part of the principles of research integrity in science.

An 'all rights reserved' licence is possible, but strongly advised against for an Open Access data set. It makes the reuse of your data very hard, if not impossible.

Access/embargo:

This is not a section in the readme file, but it is important to think about. Most datasets are archived and published under Open Access.

Restricted access can be useful if a commercial interest is involved or if the dataset contains personal data. In this case, permission is needed to access the dataset.

An embargo can be useful if your paper is not published yet, or if you wish to write another publication based on the same dataset.

Support:

More information is available on the website of the Wageningen Data Competence Center (WDCC).

If you have a question, don't hesitate to contact the Data Librarian at data@wur.nl