hvw/ 2020-03-24

Checklist for publishing data packages at Eawag

This checklist serves as a basic quality check for a data package submitted to ERIC/internal with the intention to make it available as Open Research Data in ERIC/open. The checklist is also useful for data packages that should stay internal.

- **Items that are highlighted are required for publication in ERIC/open.**
- **Items marked with a red star (*) are required for publication in ERIC/open and will be double-checked.**
- Note that some items are only "required" if applicable to the dataset, e.g. "Substances" only needs an entry if chemical substances are represented in the dataset. **If not applicable, tick the box.**
- It is your sole responsibility to consider all the points listed. Publishing data that is not well findable, is poorly documented, incomplete, or hard to re-use will impact your personal credibility as a researcher.
- **You bear the full responsibility for publication of content that violates community norms (e.g. plagiarism) and/or the law (e.g. violations of copyright or Urheberrecht).**

Package Name* (URL after ".../dataset/") :

Package DOI (if already reserved) :

1 Bibliographic metadata

DOI of the paper in case this is a publication data package and if the paper is already published. If the paper doesn't have a DOI but is in DORA, put the DORA-Id here.

* **Title**: The title of a publication data package is: "*Data for: [title of paper]*".

Authors: All persons that can claim scientific credit for the data are in the author list.

* **Authors** are given in the correct format.

Keywords: A publication data package contains at least the keywords contained in the associated paper (unless the paper covers additional topics unrelated to the data).

2 Domain specific metadata

Variables: All measured variables that pertain to the main question of the research are listed. If there are relevant variables not in the list, they have to be given in the field *Notes* at the bottom.

Substances: All chemical substances that are represented in the data are listed.

* **Substances**: Both, *scientific names* and *general terms* are given.

* **Substances'** *scientific names* are of the form
molecular formula **or** *common name* (InChI=*International Chemical Identifier*)

Example:

scientific names: atrazine (InChI=1S/C8H14ClN5/c1-4-10-7-12-6(9)13-8(14-7)11-5(2)3/h5H, 4H2,1-3H3,(H2,10,11,12,13,14))
generic terms: herbicide

Organisms: All organisms represented in the data, from which data was derived, or which are in the focus of the associated research, are listed.

* **Organisms** are listed in both, the field "*Taxa*" and the field "*Organisms*".

Systems: In case the question "Where or in which medium did the measurements or observations take place?" can be answered in an obvious way, the field "*Systems*" is not empty.

Timerange is as specific as possible.

Spatial Extent and Geographic Name(s): If the data that has one or several geographic reference(s), i.e. it was measured in the field, not in the lab,

1. "*Geographic Name(s)*" is filled out exhaustively.
2. "*Spatial Extent*" contains valid and correct GeoJSON or a GeoJSON-representation obtained from geojson.io is in the field "*Notes*" at the bottom of the form.

3 Package status & usage

* **Visibility and Status** are set to `Eawag` and (truthfully) to `complete`, respectively.

Review Level and Reviewed By are set truthfully.

The reviewer has authorized putting her name into the field *Reviewed By*.

* **Usage Contact** is set the responsible PI or group leader.

Notes contain information about meta-data that could not be properly entered into the form, such as variables measured that are not offered for selection in the field "*Variables*".

4 Resources

Resource Type is set correctly (most likely one of *Dataset*, *Text* or *Software*)

* There is a resource (file) `README.txt` or `README.md` that contains *specific metadata* (see Documentation "scientific metadata" in the *Eawag archiving guide*, https://doi.org/10.25678/000066).

All resources (files and URLs) are at least implicitly mentioned in the README-file.

* Files have **sane file-names** (see File naming conventions in the *Eawag archiving guide*, https://doi.org/10.25678/000066).

The character encoding for text files (`.txt, .csv, .md, ...`) is **UTF-8** or **ASCII**.

There are no proprietary file-formats, if can be avoided (e.g. `.doc` and `.docx` should almost always be converted to PDF).

There are **no spurious files or directories** that do not belong into the package (e.g. `.DS_Store` and `__MACOSX` in zip-archives, or `__pycache__` in Python code archives).

***.xlsx or *.ods files** are additionally saved as csv files, if possible.

* ***.xls or *.doc files** are not present. Use *.xlsx or *.docx if you really need to publish data in MS-Excel or -Word format.

CSV or TSV files (comma separated, or tab separated, *.csv, *.tsv)

have quoted text-fields, e.g. ..., "*text field value*", ...

have exactly one header row.

have units for numerical values, either in the header row, the README.txt, or a *Data Dictionary*.

are strictly rectangular.

do not have columns with duplicate header cells.

Software

is published under an OSI approved license.

is completely contained in the package. Pointers to GitHub, GitLab or similar repositories are not sufficient.

is accompanied by information about dependencies and platform compatibility.

The provenance of the data is clear: There is a reference to a publication that contains a description of the methodology, or the package itself contains that description.

Third party data is properly credited.

You have the right to publish all the content in this package.

You have considered all points above. Those that remain unchecked deviate from the recommendations on purpose.