Agroecology Partnership: Data Management Guidelines

LifeWatch ERIC

August 5, 2025

Table of contents

Introduction

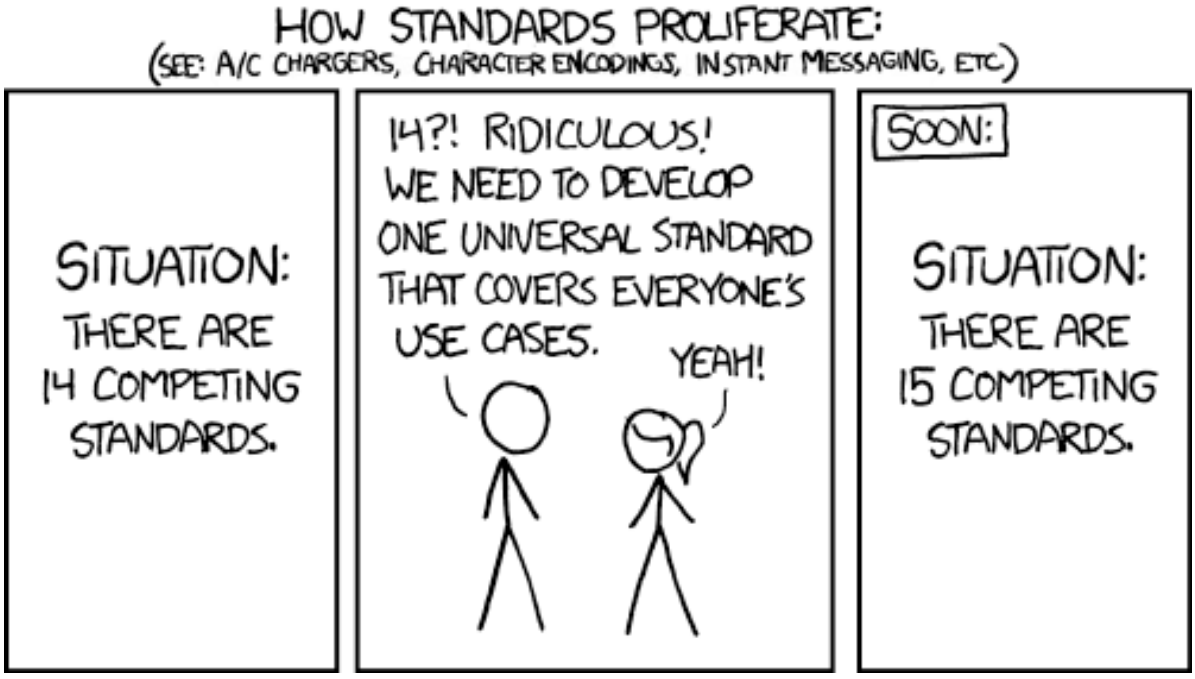
Welcome to the Data Management Guidelines documentation of the Agroecology Partnership.. This online documentation expands on the official Agroecology Partnership Data Management Plan (DMP) and complements it with more practical and technically detailed information.

While the DMP outlines the full strategy for data handling across the partnership and it is updated every 2 years, the present documentation is meant to be updated as we learn from doing. Its objective is to support partners with hands-on guidance, concrete examples, and step-by-step explanations of the data pipeline — from collection to publication — including how to make data FAIR (Findable, Accessible, Interoperable, and Reusable). This space aims to be more to the point than the DMP, helping teams implement the plan more effectively in day-to-day work. Furthermore, being a public document, we aim to offer the wider agronomy community with state-of-the-art data management practices in the Agroecology domain.

For any questions, please write us a line at agroecology-data@lifewatch.eu.

Data Standarization

Data on Agroecology are multidisciplinary: There are biological data, socio-economic, geographical and many more. The main goal is to standardize data to the preferred community standard. However, sometimes there is not a clear winning standard, or the winning standard may not meet the FAIR principles. This chapter describes the preferred data standards for each data type collected in the Agroecology Parnership. In some cases, when there is not an obvious standard, we suggest evidence-based way of structuring data that fits the FAIR principles.



Source: xkcd

Survey data

Recommendations for organizing survey data in the Agroecology partnership following a wide, spreadsheet format approach. This structure is inspired by the data structure of the American National Election Studies (ANES 2020), the recommendations of (Zimmer, Powell, and

Velásquez 2024), while applying a tidy data approach (Wickham 2014) with usability on mind and the enabling of a later transformation into Open Linked Data as structured in The Survey Ontology (Scrocca et al. 2021).

Codebook

Codebooks are files that **explain the questions** formulated in the survey. An unique identifier (a code) is assigned to each question, linking the information about the questions with the answers provided by the participants. Codebooks are typically stored as `pdf`, `docx` or `xlsx` files. Having interoperability on mind, we propose to use **text delimited files** such as `csv`. These type of files are largely used in data science. they have several advantages, including easy machine-readability and being an open format with no owner, which ensures data will remain readable and understandable by many different software for a long time.

We propose a `csv` file, delimited by semicolons `;`. We avoid using colons `,` as separators because these can be used in free text, open questions. They would affect the structure of the data. We recommend to prohibit the use of semicolons `;` in the answers provided by the participants, the questions, and in general in any use that is not deliming the columns of the table.

The codebook can be used as well during the design of the survey.

The example below shows an hypothetical survey codebook about the user satisfaction using an online platform.

`./data/codebook.csv`

Code	Label	Type	QuestionText	Values	Cardinality	QuestionType
Q1_age	Age	integer	What is your age?		1..1	SingleChoice
Q2_gender	Gender	string	What is your gender?	Female|Male|Other	1..1	SingleChoice
Q3_satisfaction	Satisfaction	integer	How satisfied are you with the platform?	1=Very unsatisfied|2=Unsatisfied|3=Neutral|4=Satisfied|5=Very satisfied	1..1	SingleChoice
Q4_improvement	Improvement	string	What would you improve in the platform?		0..n	FreeText

Each row in the codebook describes a survey question. Below is an explanation of each column:

- **Code**: A unique identifier for the question. It must be unique within the survey.
- **Label**: A short, human-readable label or name for the variable that can be used in spreadsheets or statistical software.
- **Type**: The data type of the answer. Common types include "integer", "string", "boolean", or "date".
- **QuestionText**: The full text of the question as it was asked in the survey.
- **Values**: A list of possible values for closed questions. Options are separated by vertical bars (|), and value labels can be assigned using the equals sign (=). For open or free-text questions, this field is left empty.
- **Cardinality**: Indicates how many answers are allowed. The cardinality pattern is `min..max`, where the first number is the minimum required answers and the second is the maximum allowed. `n` denotes "no fixed upper limit," so `1..1` means exactly one answer, while `0..n` means the question may be skipped or answered multiple times.
- **QuestionType**: Describes the nature of the question. Typical values include "SingleChoice", "MultipleChoice" or "FreeText"

Responses

Answers to the survey are recorded in the `./data/responses.csv` file. **Every row is the answers of a participant**, and **every column is named after the `code`** in `codebook.csv`. This allows to link easily the information about the questions without getting the responses file full of details that difficult the analysis.

`./data/responses.csv`

respondent_id	Q1_age	Q2_gender	Q3_satisfaction	Q4_improvement
001	34	female	4	I think the platform is user-friendly
002	29	male	5	Needs better support for collaboration.

We recommend to include an unique identifier for every respondent, here named as `respondent_id`. **This allows to annomymize the survey without loosing the link**

to private information about the respondents that might have been collected (e.g. name, email, address) and that it must be treated according to the European Parliament Directive 95/46/EC (General Data Protection Regulation, or GDPR). Personal data collected in the Agroecology partnership must never be published or leaked in any form.

References

ANES. 2020. "Time Series Study Full Release: User Guide and Codebook." https://electionstudies.org/wp-content/uploads/2022/02/anes_timeseries_2020_userguidecodebook_20220210.pdf.

Scrocca, Marco, Daniele Scandolari, Giulia Re Calegari, Irene Baroni, and Irene Celino. 2021. "The Survey Ontology: Packaging Survey Research as Research Objects." In *Proceedings of the 2nd Workshop on Data and Research Objects Management for Linked Open Science – Co-Located with ISWC 2021*. https://doi.org/10.4126/FRL01-006429412.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10): 1–23. https://doi.org/10.18637/jss.v059.i10.

Zimmer, Samantha A., Ryan J. Powell, and Iván C. Velásquez. 2024. *Exploring Complex Survey Data Analysis Using r: A Tidy Introduction with {Srvyr} and {Survey}*. Chapman & Hall/CRC Press. https://tidy-survey-r.github.io/tidy-survey-book/.