Data harmonisation and integration

Module 1: Introduction to data harmonisation and integration

Martina Pulieri

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 "Education and Research" - Component 2: "From research to business" - Investment 3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

Some definitions

Data harmonisation is the practice of “reconciling various types, levels and sources of data in formats that are compatible and comparable, and thus useful for better decision-making”.

Data integration is the process of combining, merging, or joining data together, in order to make what were distinct, multiple data objects, into a single, unified data object.

https://doi.org/10.1038/s41597-024-02956-3

To create an harmonised dataset or to integrate datasets, three dimensions need to be taken into consideration:

- Syntax
- Structure
- Semantics

https://doi.org/10.1038/s41597-024-02956-3

Syntax: formats

Data can come in a variety of technical formats (e.g. .csv, JSON, HTML) that can require additional processing before the data can be harmonised.

https://doi.org/10.1038/s41597-024-02956-3

Structure: conceptual schema

This refers to how different variables relate to each other within a dataset; these can vary widely across datasets. On one end of the spectrum is structured data, which are highly organised and formatted (e.g. data tables), to unstructured data with little or no fixed format (e.g. raw text, images). Different datasets can have large sources of variation not only across types of data structures but within them.

https://doi.org/10.1038/s41597-024-02956-3

Semantics: intended meaning of words

A close reading of what a given variable is intended to measure is necessary in order to properly harmonise variables across datasets. For instance, use of the same terminology does not guarantee that different datasets are measuring the same concept.

https://doi.org/10.1038/s41597-024-02956-3

Stringent vs flexible

Data harmonisation can be broadly understood as **stringent** or **flexible.**

- Stringent harmonisation refers to the use of identical measures and procedures across studies.
- Flexible harmonisation ensures that different datasets are, though not necessarily identical, inferentially equivalent and ultimately transformed into a common format.

https://doi.org/10.1038/s41597-024-02956-3

Harmonisation vs Standardisation

- Data harmonisation results in a dataset that follows a unique, cohesive ontology or taxonomy derived from conceptually similar datasets. The harmonised data itself may be constructed as a single dataset or remain dispersed across multiple datasets depending on the various ethical, legal, methodological or logistical factors at play.

- Data standardisation aims to unify data using a uniform methodology. Standardisation can be understood to be the most extreme form of stringent harmonisation possible insofar as all potential dimensions of the data (i.e. structure, syntax, semantics) are made to be identical and often held up to be the primary reference point for a given domain.

Harmonisation vs Standardisation

Aspect	Data Harmonisation	Data Integration
Main Objective	Make data consistent, compatible, and comparable	Combine data from multiple sources
Level of Transformation	Deep: standardises formats, codes, units	Shallow: merges data as-is
Handling Inconsistencies	Resolves ambiguities, converts units/names/definitions	May just aggregate without harmonising
Outcome	Uniform dataset, ready for integrated analysis	Aggregated dataset but potentially inconsistent
Typical Example	Unifying the meaning of "gender" across languages or formats (e.g. “M/F,” “Male/Female,” “1/2”)	Merging two sales tables from different branches without standardising currency or column structure

Retrospective and prospective harmonisation

Data harmonisation can take place:

- after the data has already been collected, commonly known as **retrospective harmonisation** (also known as ex-post harmonisation or output harmonisation)
- before the data has been collected, commonly known as **prospective harmonisation** (also known as ex-ante harmonisation or input harmonisation)

Retrospective data harmonisation

Retrospective harmonisation refers to harmonisation of already collected datasets.

- Original data collection is not possible and only retrospective harmonisation is feasible.
- In other cases, while original data collection may theoretically be possible, the time impermanence of primary sources may render it infeasible to fully implement.

Prospective data harmonisation

Prospective harmonisation is a distinct form of original data collection where research methodologies are harmonised before (at least some) data collection takes place.

https://doi.org/10.1038/s41597-024-02956-3

Methodologies

- **MERGING**. It entails developing a single global dataset that can encompass data across disparate datasets. The benefit of this approach is that it contains all possible information across disparate datasets but the drawback is that it may be difficult or time-consuming to develop and the resulting dataset may be unwieldy or impracticable to use.
- **MAPPING**. It creates a set of rules to relate different datasets to each other. A benefit of this approach is that original information in a given dataset can be preserved but a drawback is that mappings can become complicated if many-to-one mappings are necessary or one-to-one mappings are not possible.

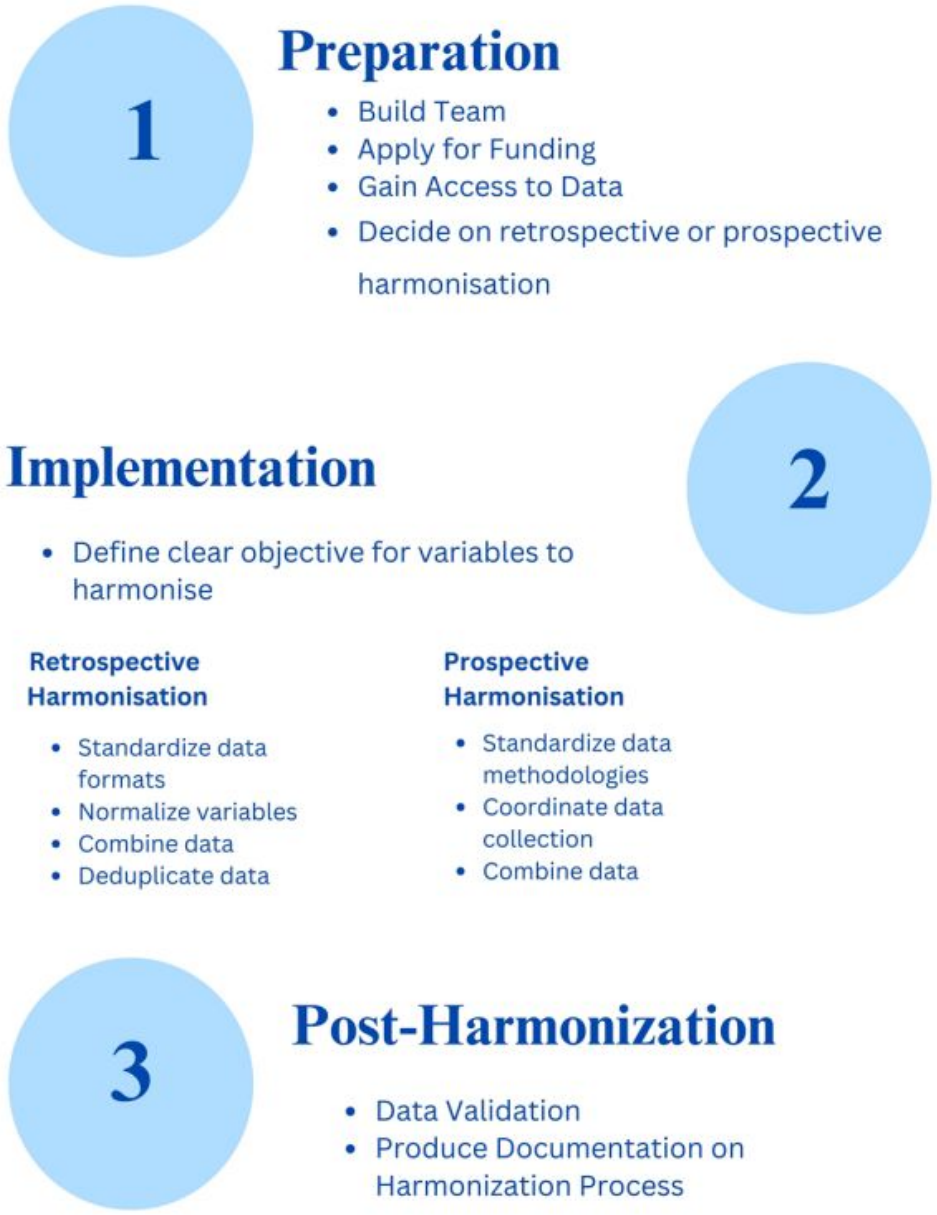


Fig. 1 General Steps for Data Harmonization.

https://doi.org/10.1038/s41597-024-02956-3

Group activity

Group 1

Question: How do coral reef fish communities respond to increasing ocean acidification and warming?

Group 2

Question: How does microplastic ingestion affect reproductive health in commercially important fish species?

Output: produce a data schema

What can be gained from data harmonisation?

1. Using harmonised data can increase the statistical power of subsequent analyses compared to those done on individual datasets;
2. Harmonising data can allow researchers to assess the generalisability of a given finding;
3. Data harmonisation can also increase the possibility of using methods which rely on increased sample size and data variation;
4. Data harmonisation may also greatly increase access to original datasets which previously were not widely accessible to the public.

https://doi.org/10.1038/s41597-024-02956-3

What can be lost from data harmonisation?

The creation of more harmonised or standardised data can come at the expense of conceptual diversity or complexity.

- Harmonising different datasets may increase the internal coherence of a given concept at the expense of minimising real and potentially important diversity in theoretical approaches toward a given topic.
- Risk that researchers may invest significant time and resources in creating standardised measures only to see them become obsolete in the face of rapid evolution in technology.

https://doi.org/10.1038/s41597-024-02956-3

Harmonisation-information tradeoff

The level of granularity in harmonising data determines the amount of information lost. Factors which affect this tradeoff include the:

1. **availability** of timely and relevant data;
2. **quality** of data collected (a particular issue when data collection methodology is not transparent or accessible);
3. **comparability**, and therefore the ultimate potential for harmonising the underlying data.

https://doi.org/10.1038/s41597-024-02956-3

What to avoid?

Combining datasets that ultimately measure different concepts, leading to false or inappropriate equivalences and nonsensical measures.

Careful exploration and comprehension of the underlying datasets to make sure they are inferentially equivalent, and thus appropriate to harmonise is imperative to avoid such outcomes.

https://doi.org/10.1038/s41597-024-02956-3

What are the limits of data harmonisation?

ERROR PROPAGATION

The data harmonisation process may propagate existing errors from original datasets or generate new ones during the data harmonisation process which can limit the validity of the subsequent data.

Tools for data harmonisation in the biomedical domain

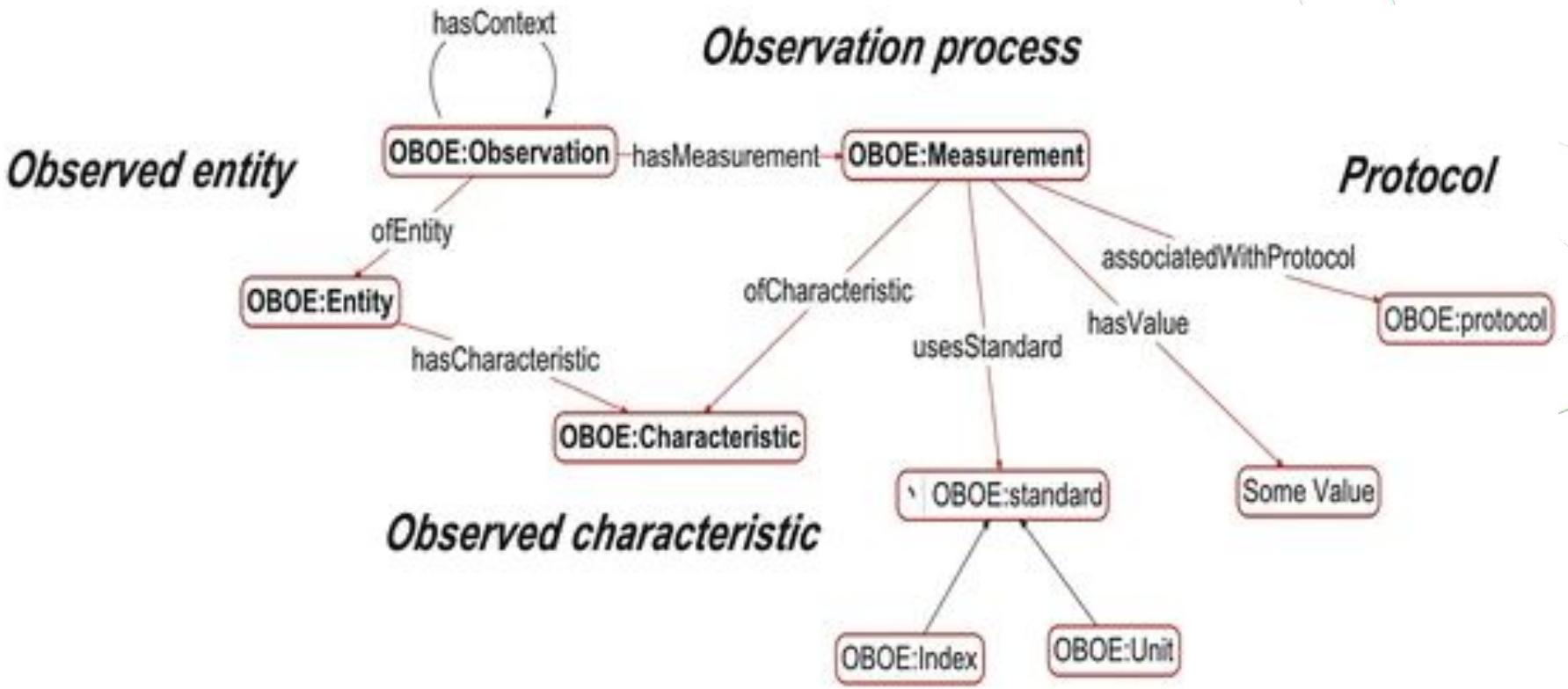
- DataHarmonizer, a standardised browser-based spreadsheet editor which is geared toward genomics data.
- HarmonizeR, an R package which makes available an algorithm can deal with missing data in omics datasets.
- Researchers, especially in epidemiology, may further benefit from making use Rmonize, an R package which provides functions to support retrospective data harmonisation, evaluation and documentation.

https://doi.org/10.1038/s41597-024-02956-3

Tools for data harmonisation for environmental sciences

- Some R packages:
  - taxize, taxonstand for taxonomic information;
  - EML for metadata schema
- Packages and platform from the Environmental Data Initiative
- LifeWatch Belgium E-lab

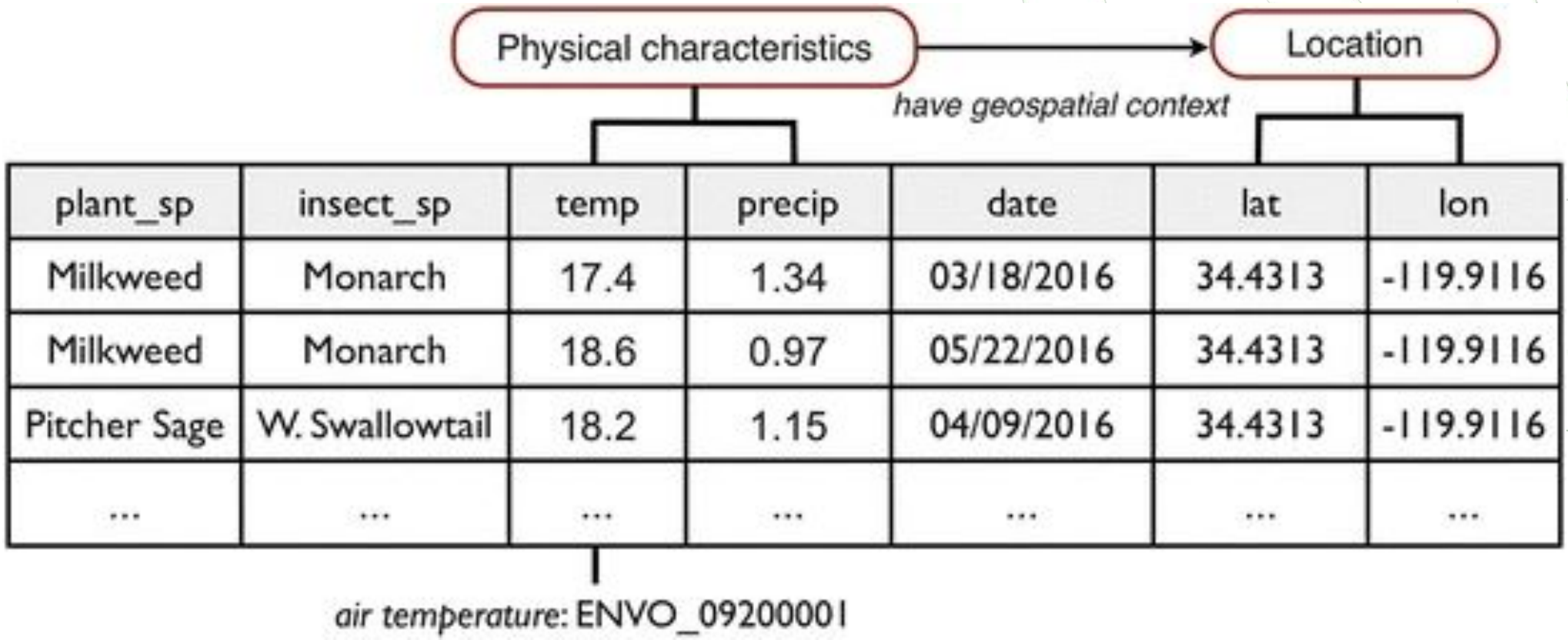
Data integration: a semantic approach



Scientific observations are decomposed into constituents representing the entity (thing or process) that is observed; the characteristics of the entity or process that were documented or measured, and assigned values; and the specific scale or units associated with those values

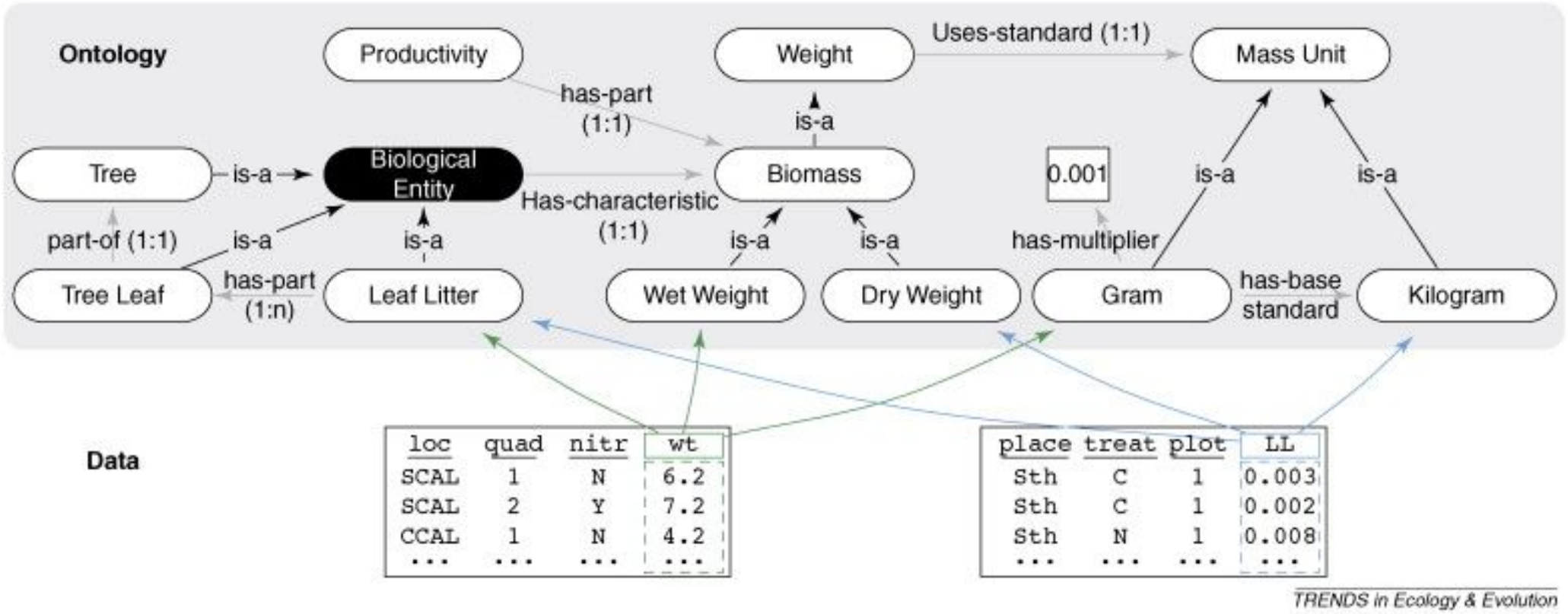
https://link.springer.com/chapter/10.1007/978-3-319-59928-1_8

Dataset mapping



https://link.springer.com/chapter/10.1007/978-3-319-59928-1_8

Dataset mapping



https://doi.org/10.1016/j.tree.2007.11.007

THANKS!

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani