

# Building a Trophic Knowledge Graph to Support Soil Food Web Reconstruction

Nicolas Le Guillarme<sup>1</sup>, Mickaël Hedde<sup>2</sup> and Wilfried Thuiller<sup>1</sup>

<sup>1</sup>Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France

<sup>2</sup>INRAE, UMR Eco & Sols, Montpellier, France

## Abstract

While food webs are pivotal tools to understand the structure, dynamics and functioning of ecosystems, their reconstruction is not trivial since feeding relationships are not always known. To this end, soil ecologists often simplify the problem by either grouping morphologically similar organisms into trophic groups with known interactions or by assuming that feeding relationships are predictable from consumer diets (e.g. frugivore or bacterivore). Interestingly, the scientific community has collected a considerable amount of information on trophic interactions and feeding habits, some of it being available in structured databases, or disseminated in the scientific and grey literature. However, the large-scale exploitation of these data for food web reconstruction is hampered by their dissemination in a multitude of heterogeneous datasets. The goal of our work is to propose a semantic data integration pipeline whose role will be to extract, aggregate, normalize, and integrate information about trophic interactions and diets into a trophic knowledge graph that will support the automatic reconstruction of soil food webs.

## Keywords

soil ecology, food web, semantic data integration, knowledge graph

## 1. Introduction

Food webs (also called trophic webs, trophic interaction networks) encode both the composition of ecological communities as well as the feeding relationships within the community. In soil ecology, food webs are often used to understand the structure and dynamics of soil assemblages, and their impact on decomposition processes and nutrient cycling [1]. Yet, reconstructing soil food webs is not a straightforward task. The nature of soil as a black-box ecosystem makes direct observations of most trophic interactions fairly impossible, and knowledge of resource preferences of many taxonomic groups of soil fauna are not well known. These preferences may be inferred from morphological similarities with species of known feeding habits or by phylogenetic proximity. The development of high-throughput species identification methods (e.g. eDNA metabarcoding), and the availability of massive amounts of data about trophic interactions and feeding habits collected by researchers over the past decades have paved the way for new knowledge-based methods to automate the reconstruction of food webs on an unprecedented scale [2]. However, despite recent efforts to facilitate access to large collections

---


*S4BioDiv 2021: 3<sup>rd</sup> International Workshop on Semantics for Biodiversity, held at JWOW 2021: Episode VII The Bolzano Summer of Knowledge, September 11–18, 2021, Bolzano, Italy*

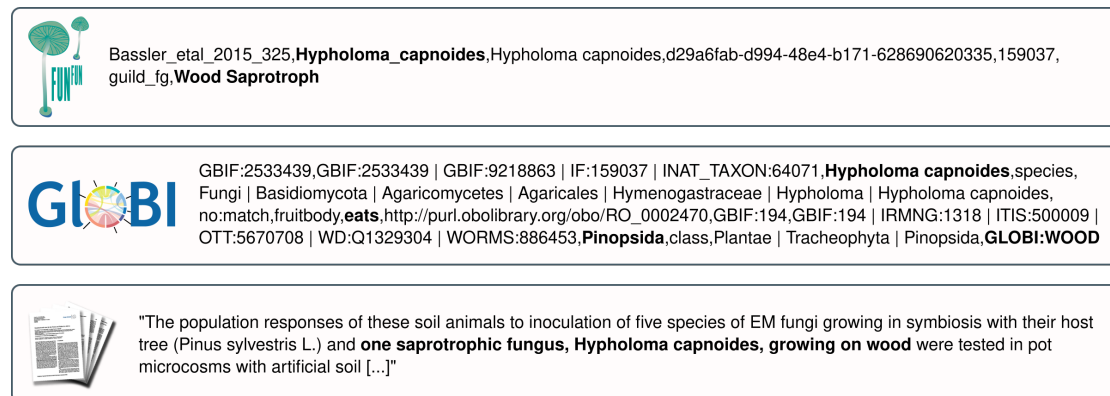
✉ nicolas.leguillarme@univ-grenoble-alpes.fr (N. Le Guillarme); mickael.hedde@inrae.fr (M. Hedde)

ORCID 0000-0003-4559-7579 (N. Le Guillarme); 0000-0002-6733-3622 (M. Hedde); 0000-0002-5388-5274 (W. Thuiller)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** The same information about the feeding habits of saprotrophic fungus *Hypholoma capnoides* as it can be found in fungaltraits (a fungal trait database), GloBI (the Global Biotic Interactions database), and an article published in *Soil Biology & Biochemistry*. Both fungaltraits and GloBI are structured databases, while textual data is by nature unstructured. Fungaltraits focuses on trophic groups (*Wood Saprotroph*), while GloBI is a trophic interaction database, using terms from the RO ontology (e.g. *eats*) to describe the nature of the interaction. GloBI provides unambiguous species identification using taxon identifiers (e.g. GBIF:2533439), whereas in the other two data sources the consumer is referred to by its scientific name only.

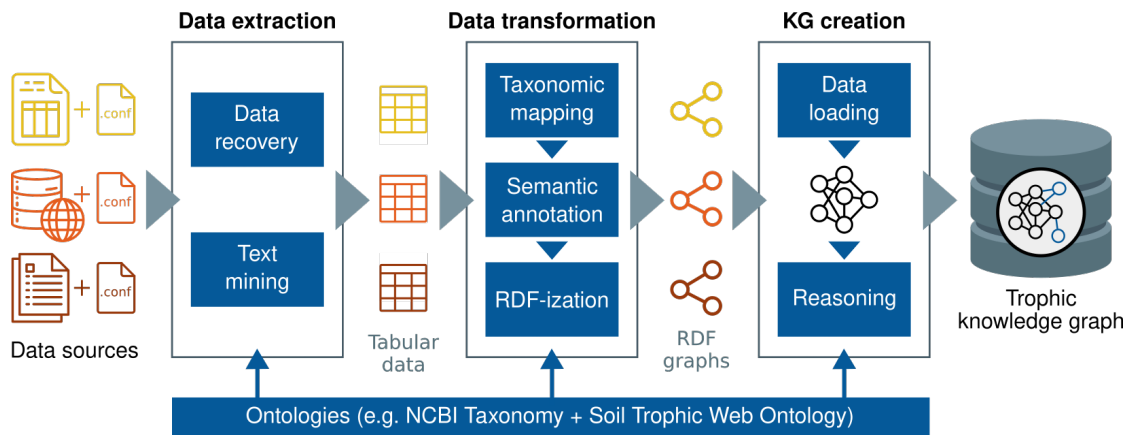
of structured data on species interactions [3], information about trophic interactions and diets (e.g. fungivory, xylophagy...) is still largely scattered in multiple heterogeneous datasets, as well as in the scientific and grey literature. These datasets may have different levels of structuredness and do not usually share common taxonomies and trophic classifications (Fig. 1), which hampers data interoperability and reuse and impose limits on the scale at which studies can be carried out.

To address this issue, we are developing a semantic data integration pipeline whose role will be to integrate existing trophic datasets with a reference taxonomy and a new ontology of soil trophic webs. The resulting knowledge graph will offer unified access services over a large collection of trophic information from heterogeneous datasets. In addition, its formal semantics will allow additional knowledge about trophic group membership to be deduced from trophic interaction data, and vice versa, thus supporting food web reconstruction.

## 2. Semantic Integration of Trophic Data

Our semantic integration pipeline receives data from multiple sources in different formats: structured data from open-access relational/graph databases or in-house datasets stored as tabular files, semi-structured data (XML, JSON), and ultimately, unstructured data as free-form text from scientific papers or wiki pages. The architecture of the pipeline is shown in Fig. 2. This architecture comprises three main building blocks:

**Data extraction.** The role of this module is to fetch data from their respective sources and to store the extracted data as structured datasets in a local directory. The data extraction



**Figure 2:** Our ontology-based data integration pipeline adopts an Extract-Transform-Load approach to build a trophic knowledge graph from heterogeneous data sources.

module currently provides methods to access local or remote files (e.g. download dataset dumps from the Web), and to access RDF data via SPARQL endpoints. This module will also rely on text mining techniques (e.g. trophic relation extraction) to transform unstructured textual sources (scientific publications, wiki pages...) into structured datasets.

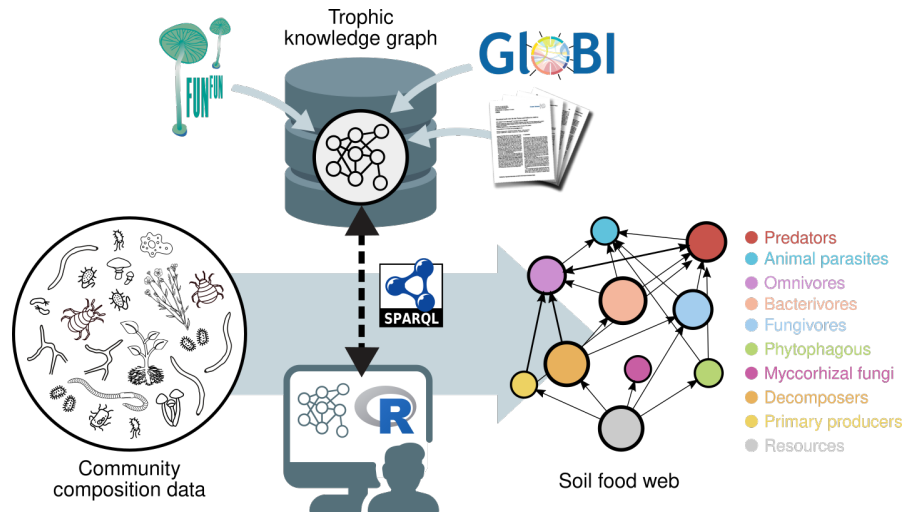
**Data transformation.** The data transformation module is responsible for the normalization and RDF-ization of the extracted data. In particular, all taxonomic data (taxon names or identifiers) are mapped to entities in the NCBI Taxonomy using *nomer* [4], an open-source taxonomic alignment software. The NCBI Taxonomy was chosen as the reference taxonomy because it is the only taxonomy available in the form of an OWL ontology, which allows reasoning on taxonomic knowledge.

Information concerning trophic interactions (consumer-resource relationships) and group membership is represented using terms from the Soil Trophic Web Ontology (STWO, [5]). This OWL ontology provides a formal description of trophic knowledge, including logical definitions of trophic groups that allow new facts to be deduced using OWL reasoning. STWO leverages existing resources by importing modules from existing OBO ontologies (e.g. RO, ECOCORE, ENVO...) and adds new classes for missing groups and resources.

Semantically annotated data are then transformed into RDF triples using *RMLMapper* [6], an open-source Java library that executes RML mapping rules to generate RDF data from various input formats (e.g. CSV, JSON, XML, RDF...). These rules are dataset-specific, which means that a new set of rules have to be written manually for each new data source.

At the end of the data transformation step, all datasets have been converted into named RDF graphs, the name of the graph being used to keep track of data provenance.

**Knowledge graph creation.** This module simply loads all the datasets in N-Quads format into a triplestore. The result is a knowledge graph where taxonomic entities (consumers) are linked to resources (other taxonomic entities, anatomical entities, environmental mate-



**Figure 3:** The trophic knowledge graph built using our semantic data integration pipeline provides a single access point to trophic group and interaction data from a multitude of sources, thus supporting the reconstruction of food webs from community composition data. An R package allows easy access to the knowledge contained in the graph by encapsulating SPARQL queries.

rial...) by trophic interactions, either explicitly (e.g. the triple `ncbi:CarabusHispanicus ro:eats ncbi:Gastropoda`), or implicitly through trophic group information (e.g. the triple `ncbi:CarabusHispanicus rdf:type stwo:malacophage`).

Additional facts about trophic group membership or potential trophic interactions can be deduced from existing knowledge using OWL reasoning. For instance, the fact that *C. hispanicus* is a malacophagous organism can be inferred from the following facts: *C. hispanicus* consumes gastropods, Gastropoda is a taxonomic class within the phylum Mollusca, and the class of malacophagous organisms is logically defined in STWO as the set of all organisms consuming molluscs. Similarly, knowing that *C. hispanicus* is malacophage, it is possible to deduce that *C. hispanicus* interacts trophically with members of the phylum Mollusca.

This type of reasoning can be carried out by the triplestore's built-in reasoner (if any) or delegated to an external reasoner (e.g. Pellet, HermiT...).

Our implementation relies on generic Python components that are instantiated and assembled into a dedicated pipeline for each new data source by providing a set of configuration files (including RML mapping rules for this specific dataset). External tools (e.g. nomer, RMLMapper) are run using Docker. Pipelines are scheduled and monitored using Apache Airflow. The entire integration workflow is fully automated and can be configured so that the knowledge graph is rebuilt on a regular basis (weekly, monthly...) to account for source dataset updates. Finally, end-users (mainly soil ecologists) can interact with the knowledge graph via an R package that implements the most common SPARQL queries, thus providing a friendly interface to users that do not have SPARQL knowledge (Fig 3).

### 3. Future Work

Our vision is that of a semantic data integration pipeline that will enable ecologists to make the most of available information about soil trophic ecology, regardless of the initial format and location of this information. To achieve this, we are working on three fronts: the development of a domain ontology to represent knowledge about soil trophic ecology, the implementation of a user-friendly trophic knowledge graph construction pipeline, and last but not least, the development of trophic information extraction tools to handle unstructured data sources.

We implemented a proof-of-concept of our data integration pipeline that is fully functional, easily extensible to new data sources, and generic enough so that it could be used to integrate other types of knowledge (e.g. organism traits, habitats, etc.), provided that the appropriate ontological resources exist. We are working on making the pipeline even more accessible to non-expert users. For instance, we would like to provide a easier way to generate RML mapping rules for new data sources. In a near future, we plan to focus on quality assessment by improving data provenance tracking and developing error detection methods.

Trophic data integration at scale still faces a number of challenges: the existence of several reference taxonomies whose mapping to the NCBI Taxonomy (currently the only taxonomy available in ontology form) is far from trivial (e.g. the SILVA taxonomy), the limited lifetime of organism-related information due to taxonomic instability, the challenge of large-scale ontological reasoning... These are just a few examples of the many opportunities for collaboration between ecologists and knowledge engineers that could benefit biodiversity science.

### Acknowledgments

The research received funding from the French Agence Nationale de la Recherche (ANR) through the GlobNets (ANR-16-CE02-0009) project and through MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

### References

- [1] S. Scheu, The soil food web: structure and perspectives, *European journal of soil biology* 38 (2002) 11–20.
- [2] Z. G. Compson, W. A. Monk, C. J. Curry, D. Gravel, A. Bush, C. J. Baker, M. S. Al Manir, A. Riazanov, M. Hajibabaei, S. Shokralla, et al., Linking DNA metabarcoding and text mining to create network-based biomonitoring tools: A case study on boreal wetland macroinvertebrate communities, *Advances in ecological research* 59 (2018) 33–74.
- [3] J. H. Poelen, J. D. Simons, C. J. Mungall, Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets, *Ecological Informatics* 24 (2014) 148–159.
- [4] J. Poelen, globalbioticinteractions/nomer, 2021. URL: <https://doi.org/10.5281/zenodo.4925111>.
- [5] N. Le Guillarme, M. Hedde, W. Thuiller, STWO : an ontology for soil food web reconstruction, S4BioDiv 2021: 3rd International Workshop on Semantics for Biodiversity (2021).
- [6] RML.io, RMLio/rmlmapper-java, 2021. URL: <https://github.com/RMLio/rmlmapper-java>.