Séminaire rdatadev
5 mars 2024

LabEx OSUG
Observatoire des Sciences de l'Univers de Grenoble

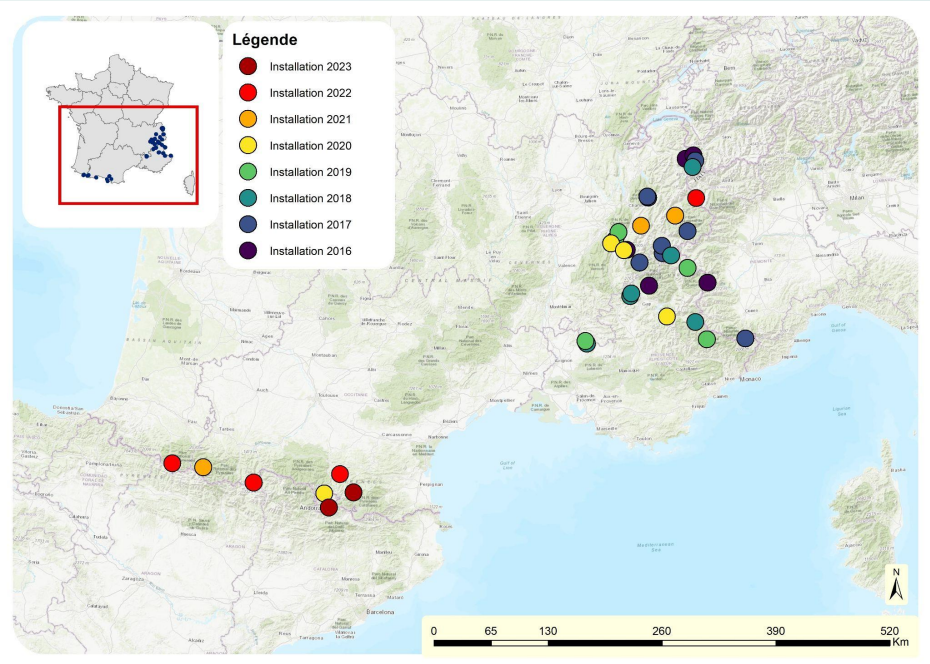
Intégration sémantique de données de biodiversité et construction d'un graphe de connaissances en écologie

Nicolas Le Guillarme
Laboratoire d'Écologie Alpine, Grenoble, France

LECA Laboratoire d'Ecologie Alpine

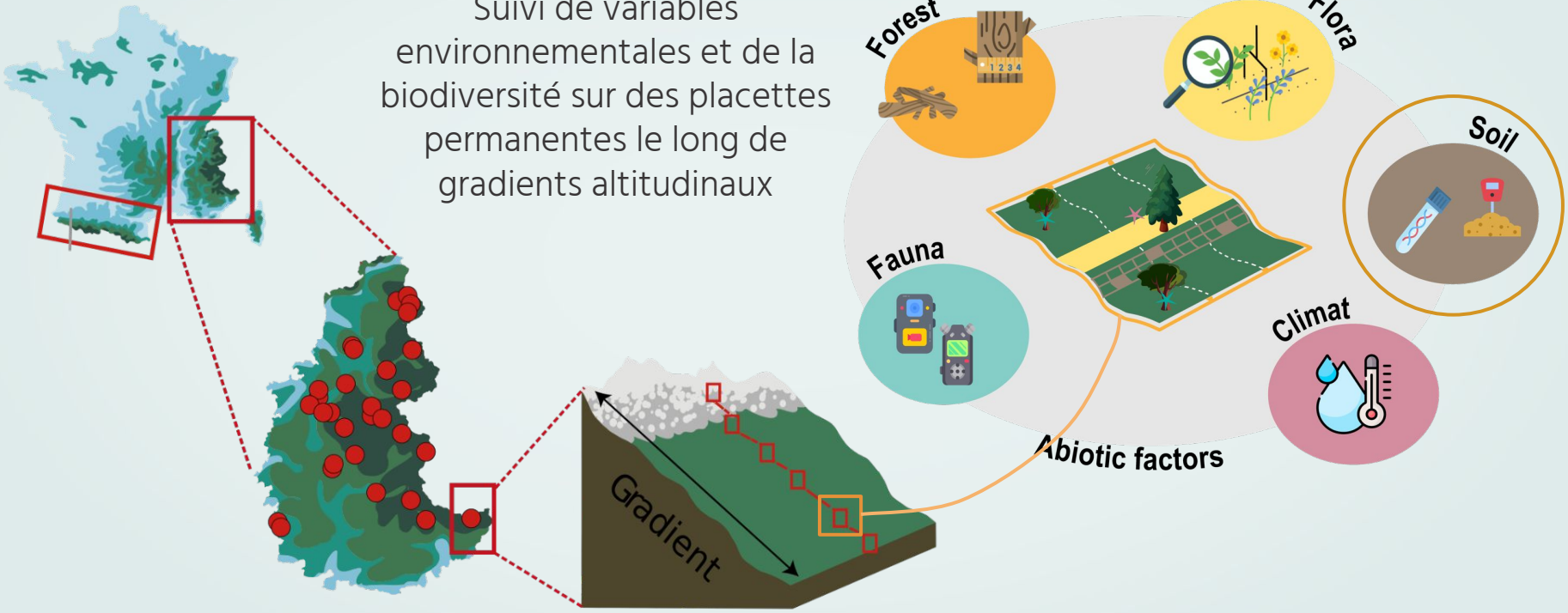
L'observatoire ORCHAMP

Observer, étudier et modéliser la biodiversité des territoires de montagne pour c**omprendre et prédire** les répercussions des changements environnementaux sur la biodiversité et le fonctionnement des écosystèmes dans les Alpes françaises et les Pyrénées



L'observatoire ORCHAMP

Suivi de variables environnementales et de la biodiversité sur des placettes permanentes le long de gradients altitudinaux



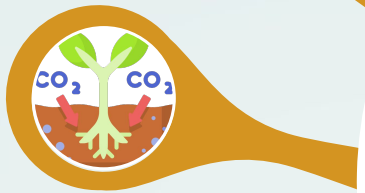
Thuiller, W., Saillard, A., Abdulhak, S., Augé, V., Birk, C., Bonet, R., Choler, P., Delestrade, A., Kunstler, G., ... & The Orchamp Consortium. (2024). ORCHAMP: Monitoring Biodiversity and Ecosystem Functioning Across Space and Time in Mountainous Regions. *Manuscript in preparation.*

La biodiversité du sol représente près de 26 % des espèces vivantes connues de la planète

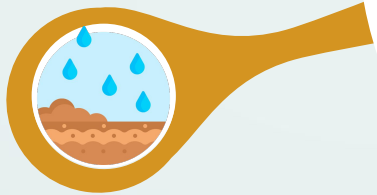
(Estimation en 2021)

Recyclage de la matière organique, croissance des plantes

Stockage du carbone, régulation du climat

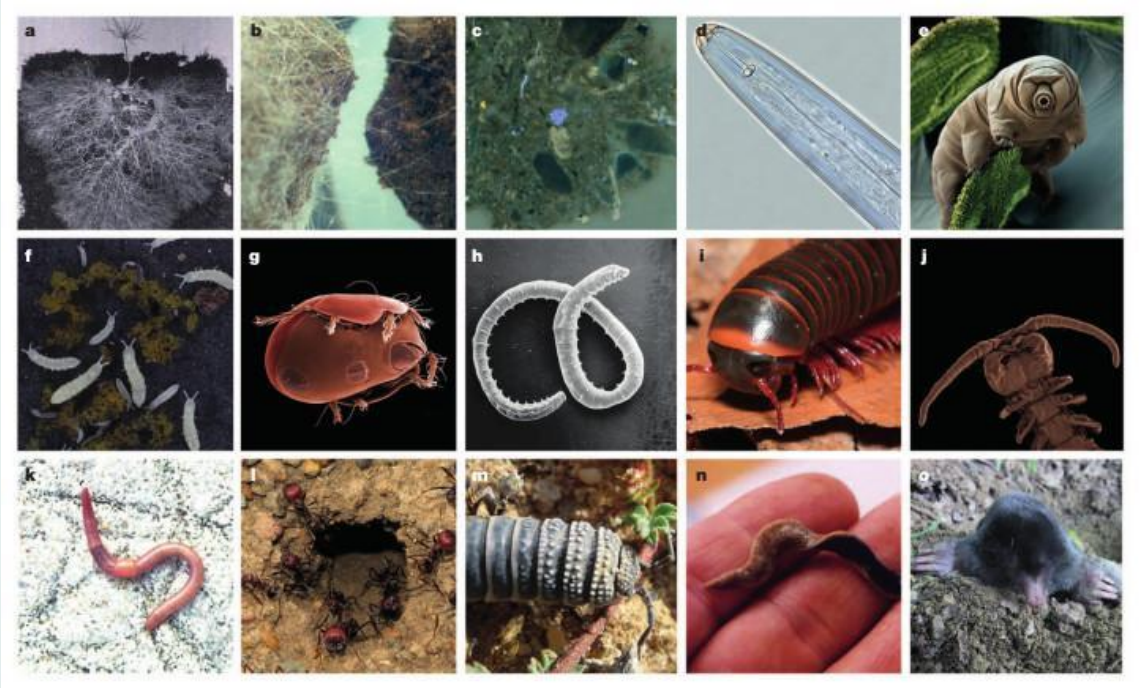


Régulation du cycle de l'eau, purification



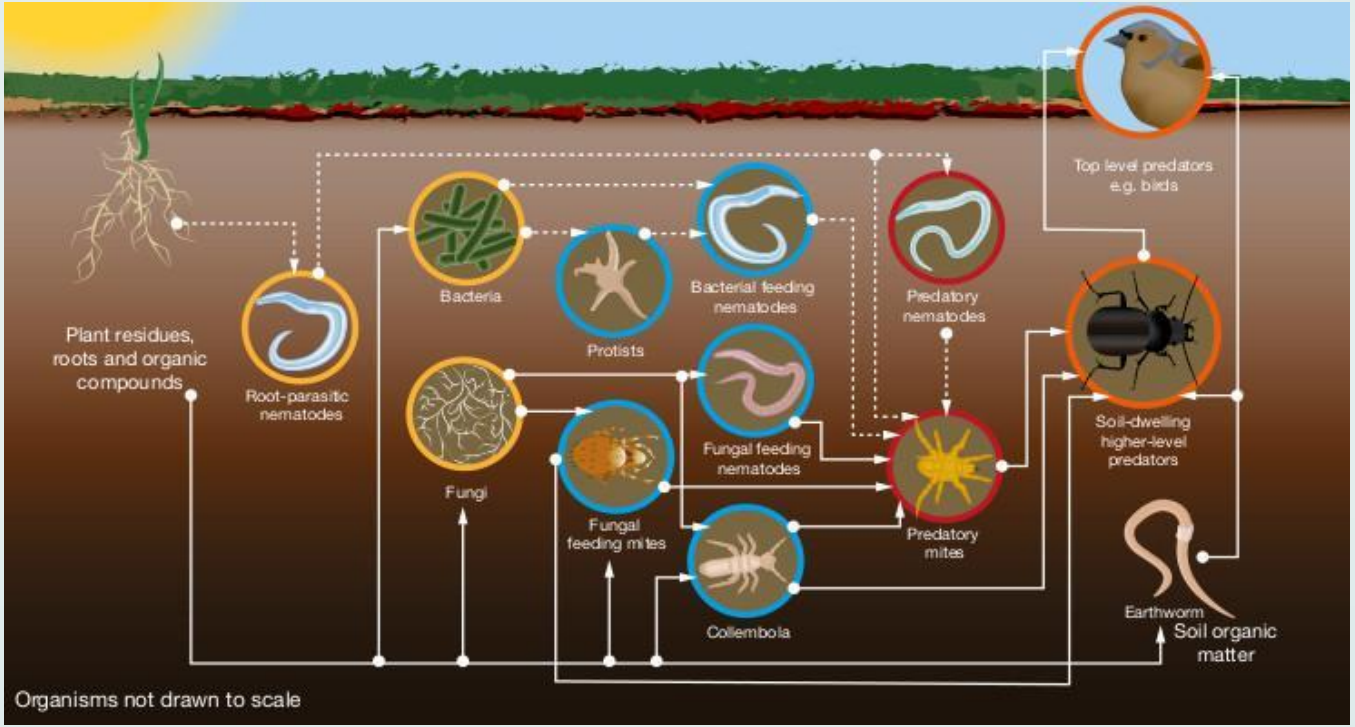
Production alimentaire, santé humaine

Un écosystème multitrophique complexe

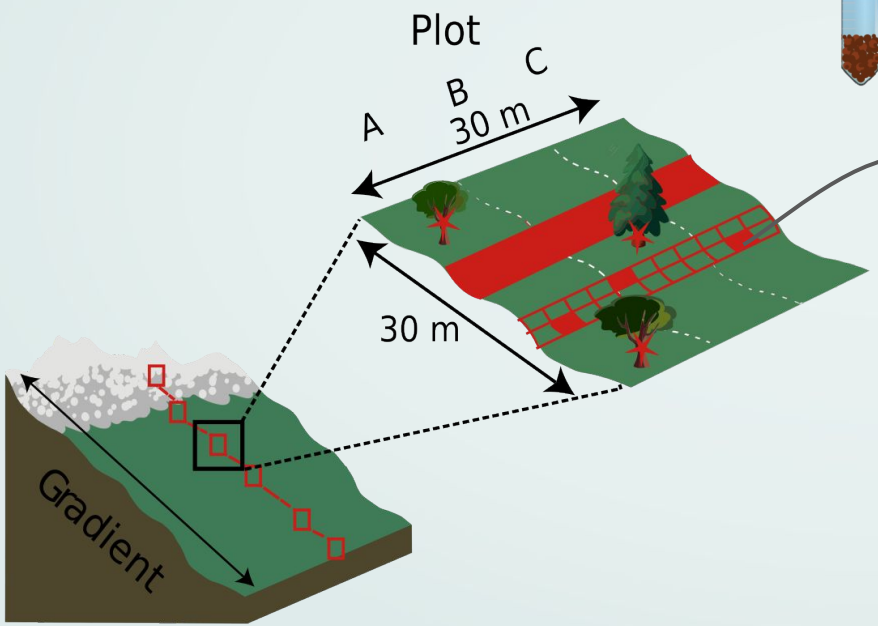


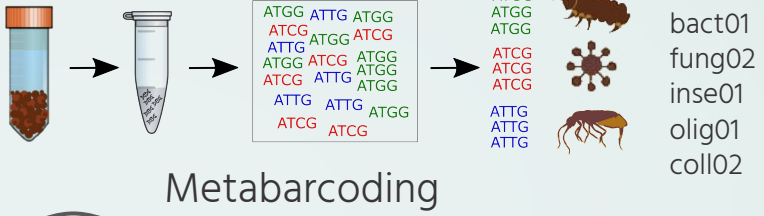
Bardgett & van der Putten, *Nature* (2014)

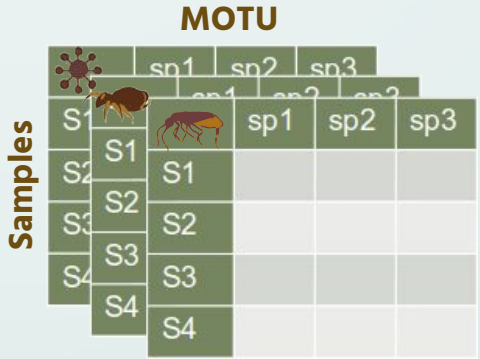
Un écosystème multitrophique complexe

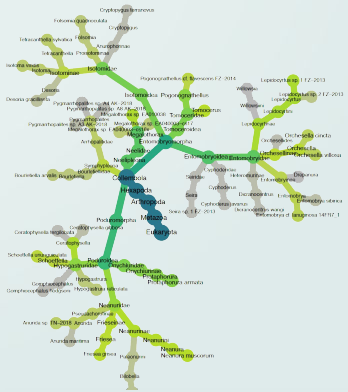


Modélisation des réseaux trophiques du sol

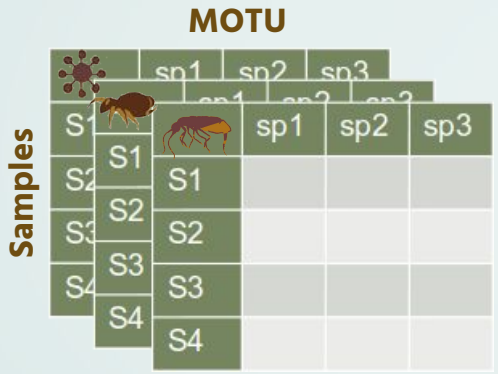




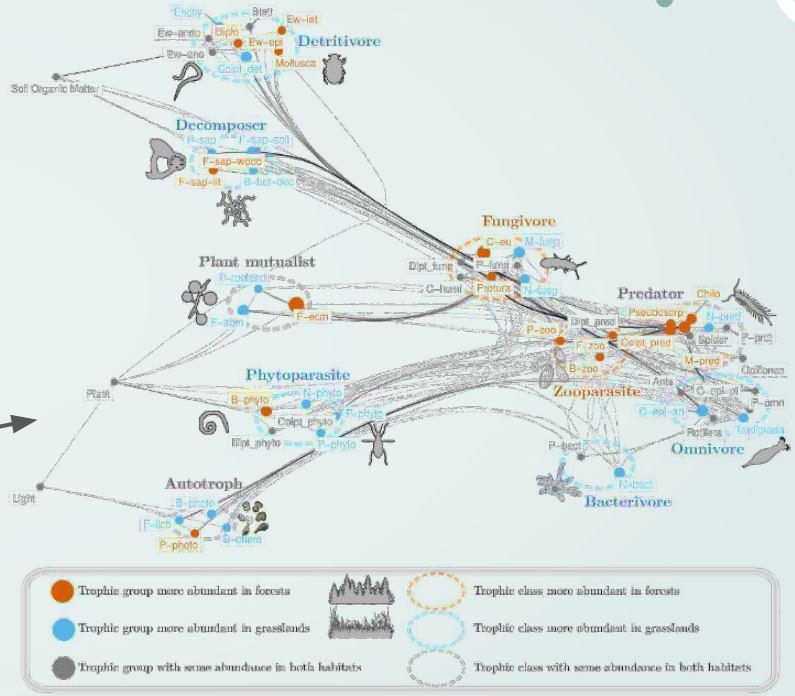




Modélisation des réseaux trophiques du sol



Bases de données trophiques



Calderón-Sanou, I., Ohlmann, M., Münkemüller, T., Zinger, L., Hedde, M., Lionnet, C., ... & Orchamp Consortium. (2024). Mountain soil multitrophic networks shaped by the interplay between habitat and pedoclimatic conditions. Soil Biology and Biochemistry, 190, 109282.

Des sources d'informations distribuées et hétérogènes

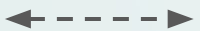
Base de traits fonctionnels pour la fonge mondiale téléchargeable au format XLSX

Base de données d'interactions écologiques en ligne accessible par une API

Base de traits pour les invertébrés du sol téléchargeable au format CSV

De nombreuses bases de données collectées et stockées en local par différentes équipes (*long-tail data*)

Des sources d'informations distribuées et hétérogènes



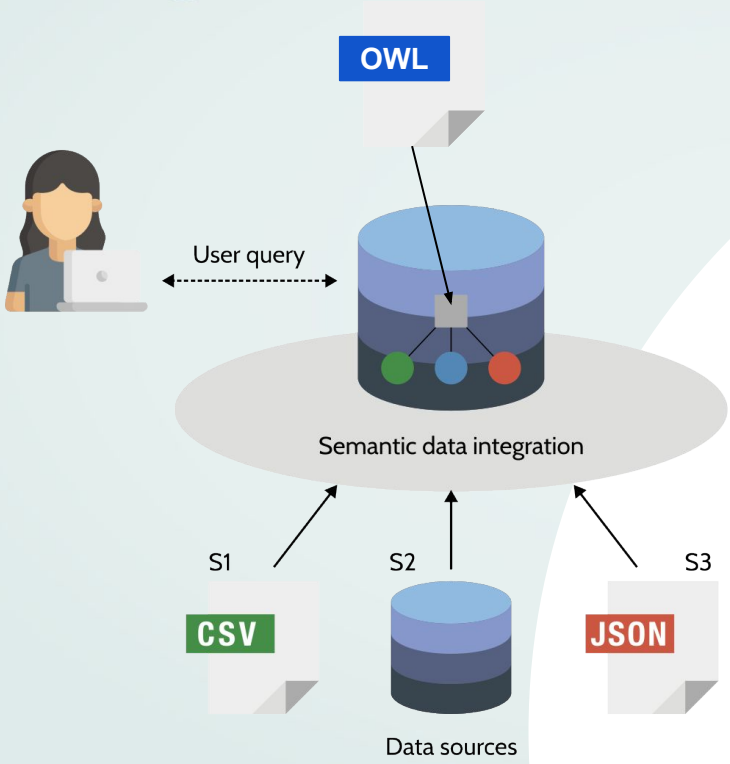
Intégration sémantique de données

GloBI

CSV

BEISI

Intégration sémantique de données



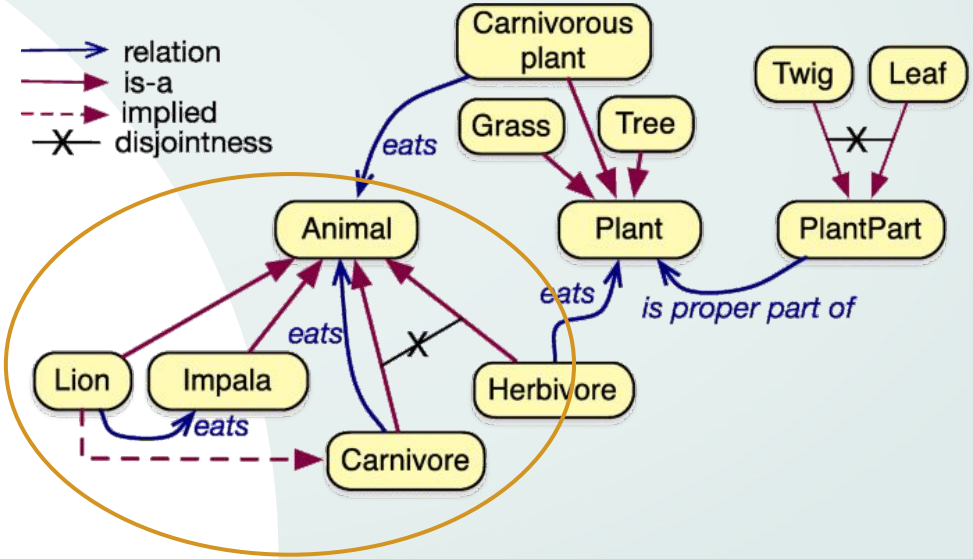
Intégration → combinaison de données issues de **sources hétérogènes** au sein d'une base de connaissance unique.

Sémantique → utilisation d'une **ontologie** pour réconcilier les différences dans l'interprétation de la "signification" des données.

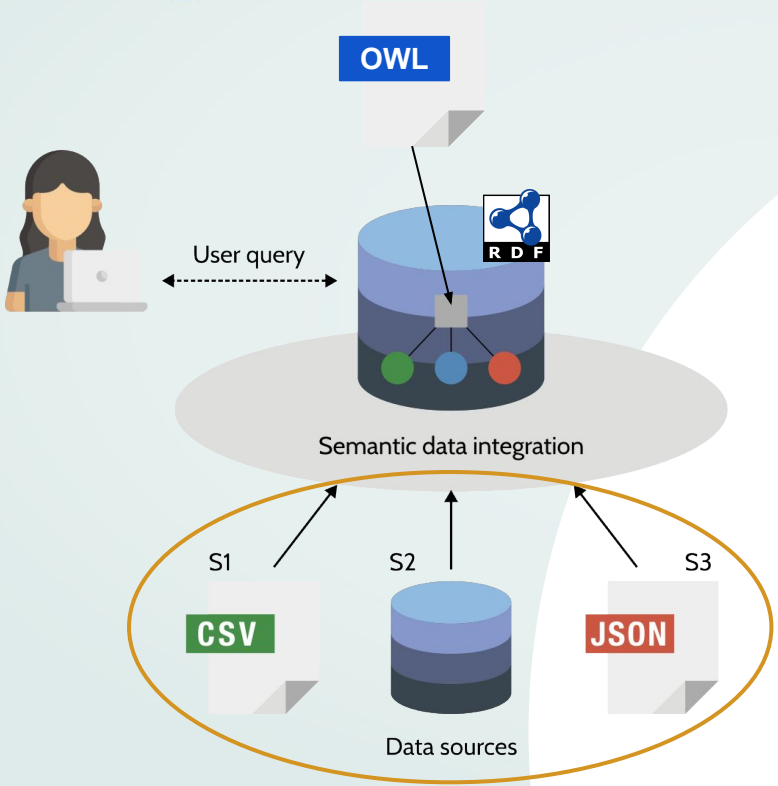
Ontologie pour la représentation des connaissances

Ontologie = un **modèle formel de la connaissance** dans un domaine qui décrit les concepts et leurs relations à l'aide de la logique mathématique.

De nouvelles connaissances (implicites) peuvent être dérivées grâce au **raisonnement automatique**.



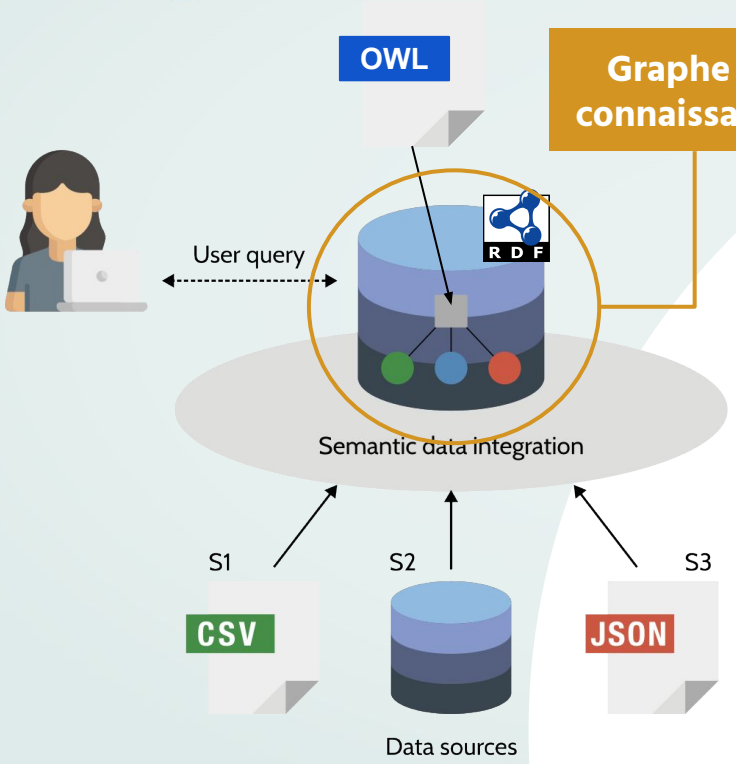
Le format RDF



Les hétérogénéités schématiques (structurelles) sont résolues en transformant les données dans un format commun : **le format RDF** (Resource Description Framework).

Le langage de représentation des connaissances OWL est construit sur le modèle de données RDF.

Graphe de connaissance

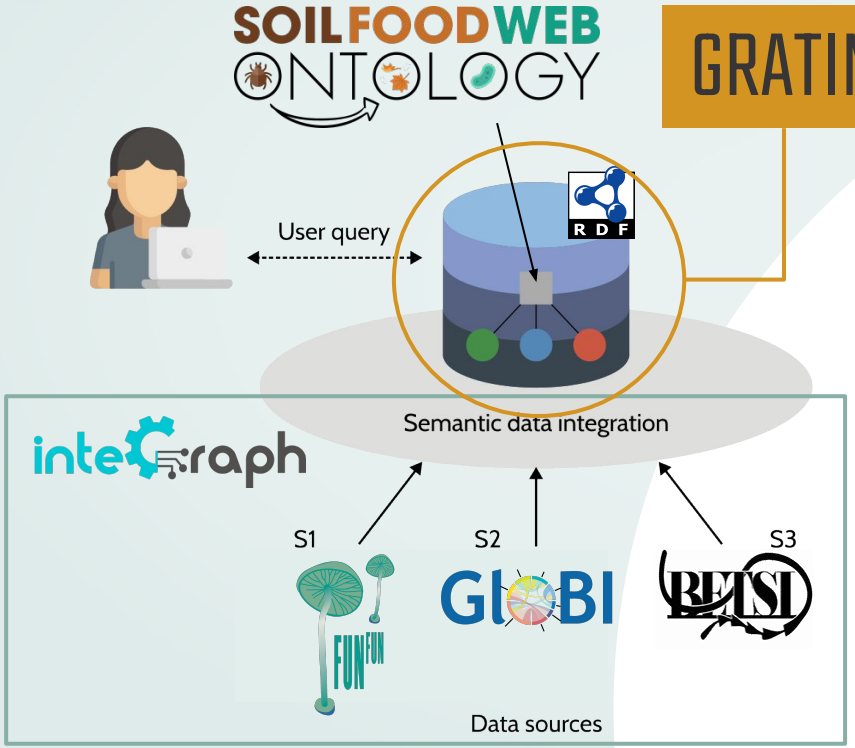


Graphe de connaissance = une base de données RDF dont la sémantique est controlée par une ontologie.

> "A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge."
>
> [Ehrlinger and Wöß, 2016]

Intégration sémantique de données = la construction d'un graphe de connaissances à partir de sources de données hétérogènes.

Construction d'un graphe de connaissance en écologie trophique du sol



La **Soil Food Web Ontology** est une ontologie OWL qui décrit formellement les concepts liés aux réseaux trophiques du sol.

inteGraph est une boîte à outil permettant de générer et de contrôler l'exécution de pipelines de sémantification de données de biodiversité.

GRATIN est un graphe de connaissance fournissant un accès unifié à des informations multisources, multitaxons, multitrophiques.

SOILFOODWEB ONTOLOGY

https://soilfoodwebontology.github.io/

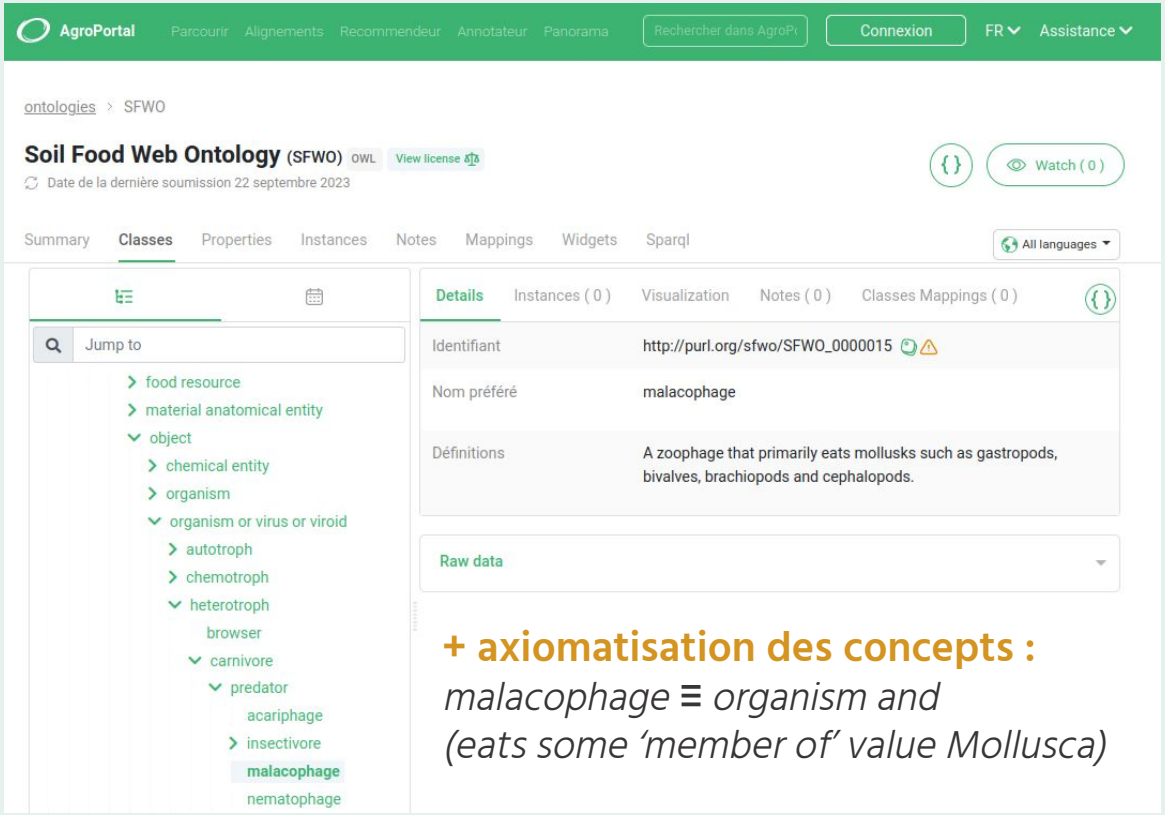
20+ contributeurs

500+ concepts

1000+ termes

160+ groupes trophiques pour la classification de la faune du sol et de la fonge

Un développement continu ouvert à tous



+ axiomatisation des concepts :

malacophage ≡ organism and (eats some 'member of' value Mollusca)

Le Guillarme, N., Hedde, M., Potapov, A. M., Martínez-Muñoz, C. A., Berg, M. P., Briones, M. J., ... & Thuiller, W. (2023). The soil food web ontology: Aligning trophic groups, processes, resources, and dietary traits to support food-web research. *Ecological Informatics*, 102360.

SOILFOODWEB ONTOLOGY

https://soilfoodwebontology.github.io/

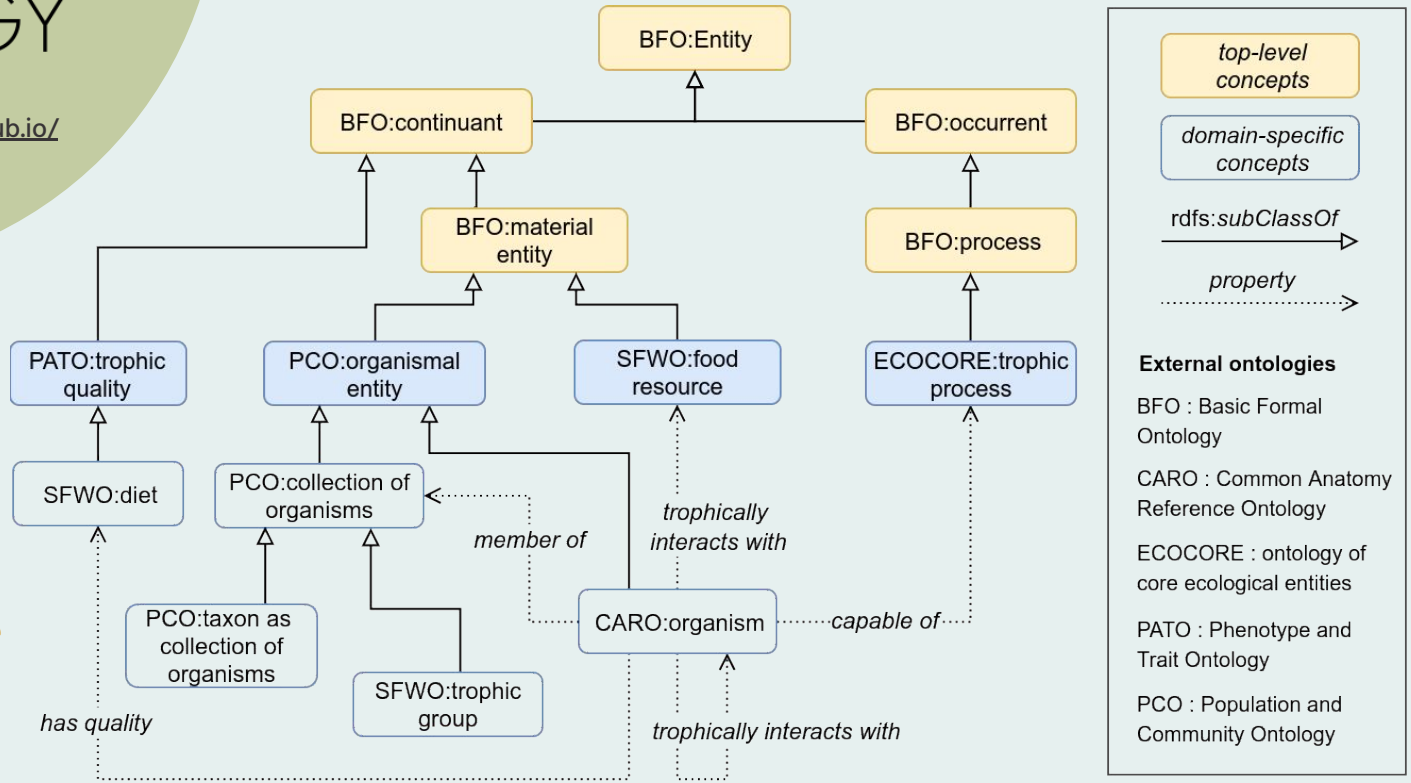
20+ contributeurs

500+ concepts

1000+ termes

160+ groupes trophiques pour la classification de la faune du sol et de la fonge

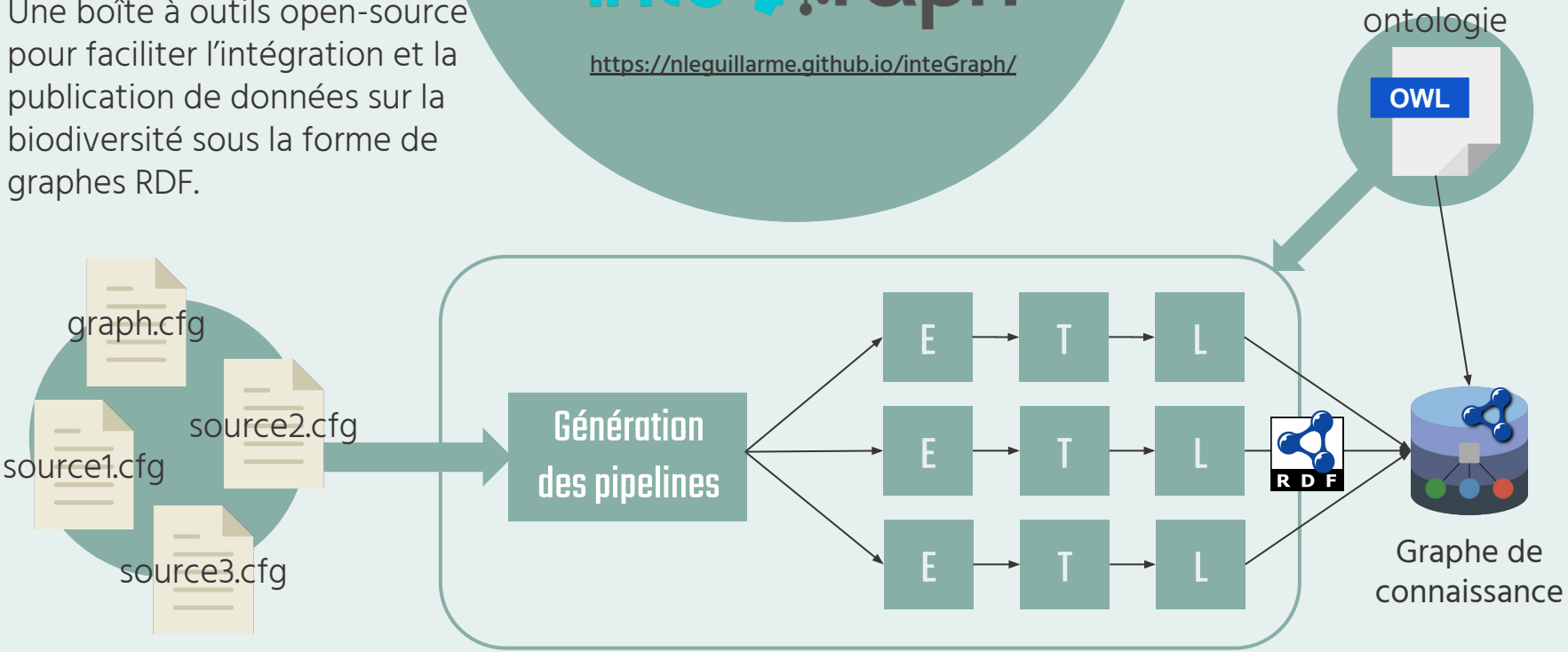
Un développement continu ouvert à tous



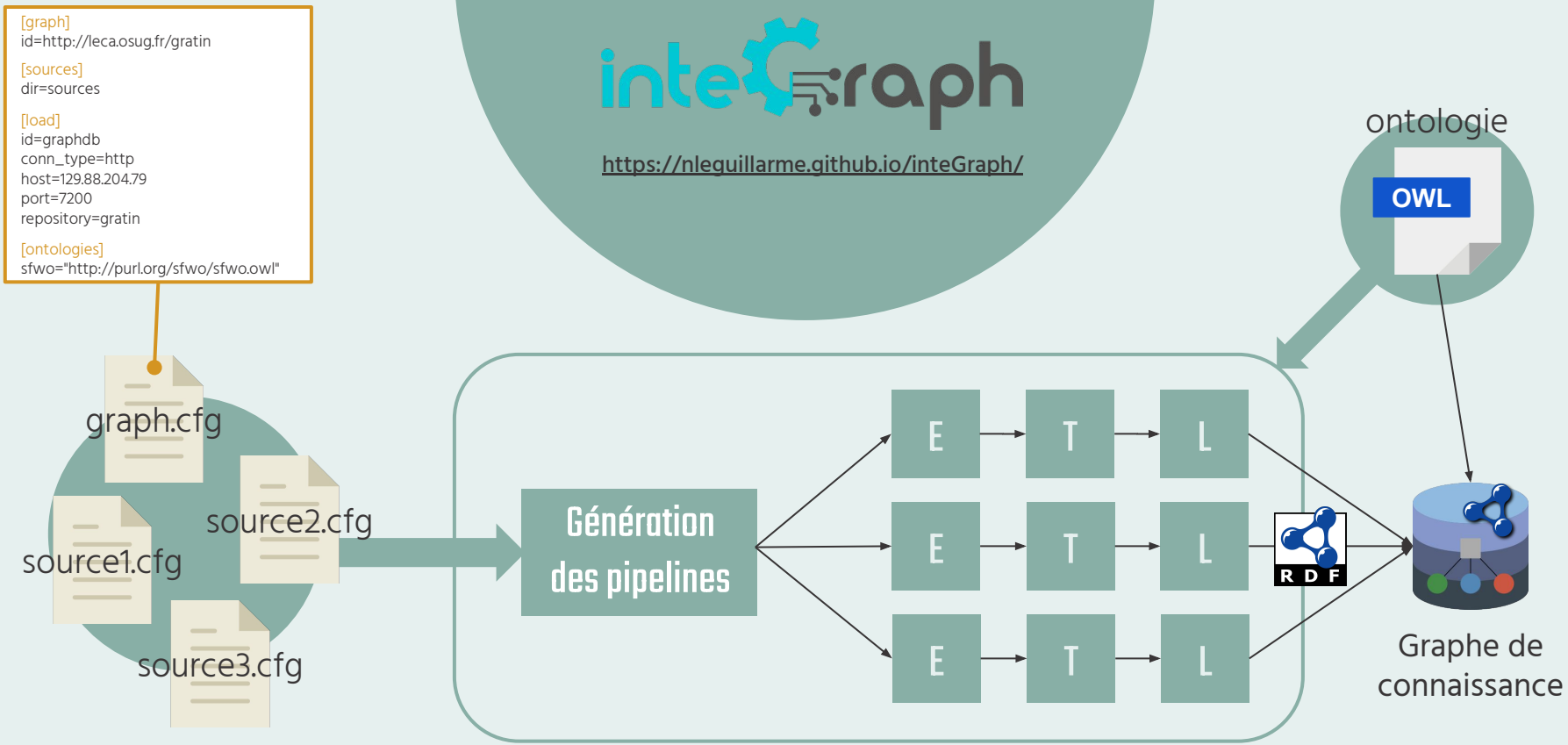
Le Guillarme, N., Hedde, M., Potapov, A. M., Martínez-Muñoz, C. A., Berg, M. P., Briones, M. J., ... & Thuiller, W. (2023). The soil food web ontology: Aligning trophic groups, processes, resources, and dietary traits to support food-web research. *Ecological Informatics*, 102360.

Une boîte à outils open-source pour faciliter l'intégration et la publication de données sur la biodiversité sous la forme de graphes RDF.

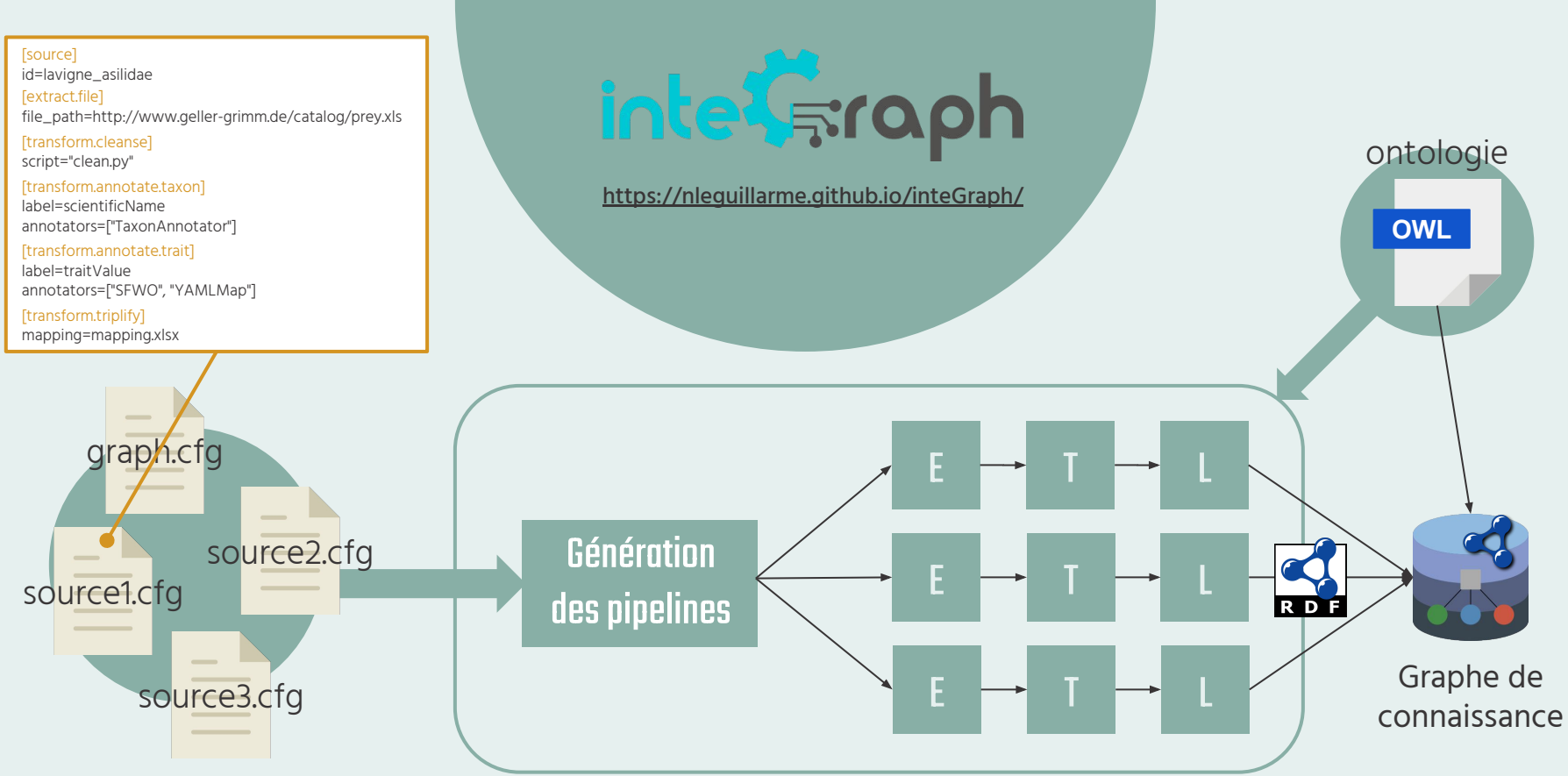
https://nleguillarme.github.io/inteGraph/



Le Guillarme, N., & Thuiller, W. (2023). A practical approach to constructing a knowledge graph for soil ecological research. *European Journal of Soil Biology, 117*, 103497.

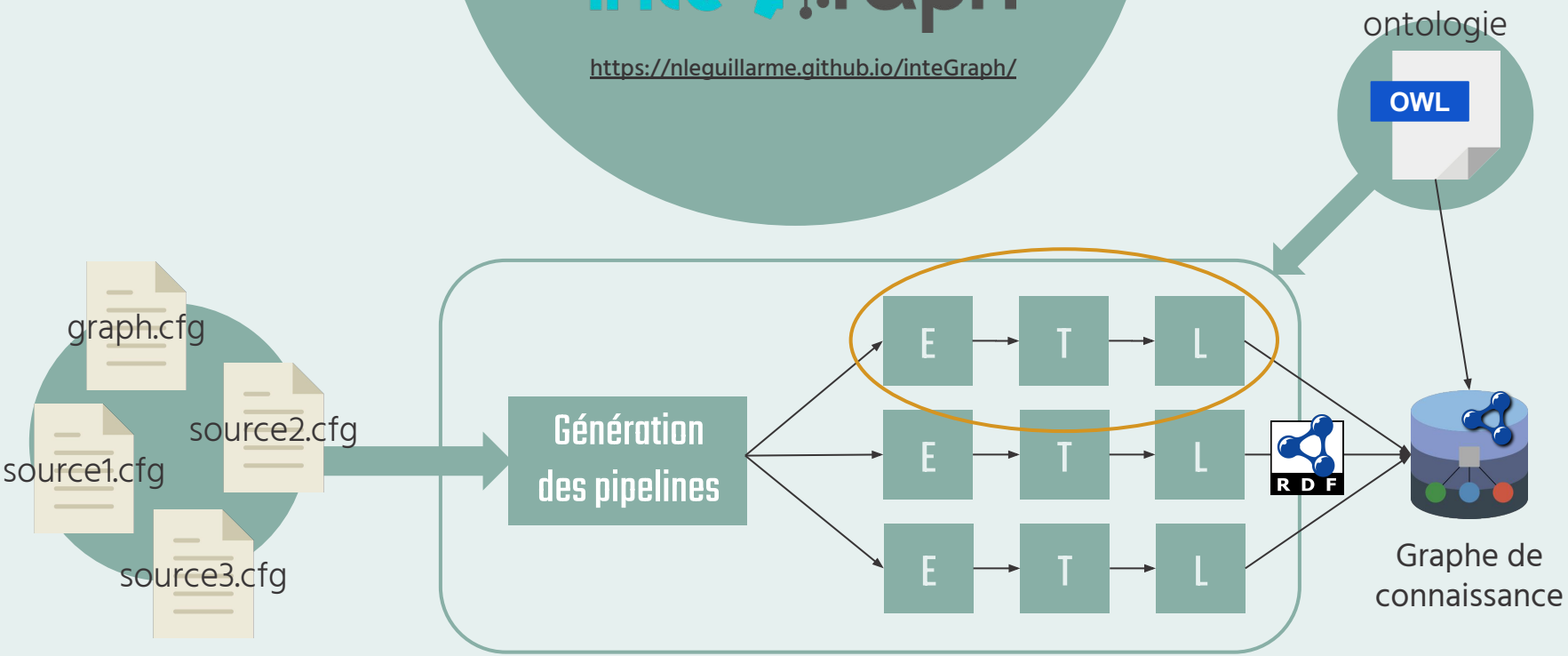


Le Guillarme, N., & Thuiller, W. (2023). A practical approach to constructing a knowledge graph for soil ecological research. *European Journal of Soil Biology, 117*, 103497.



Le Guillarme, N., & Thuiller, W. (2023). A practical approach to constructing a knowledge graph for soil ecological research. *European Journal of Soil Biology, 117*, 103497.

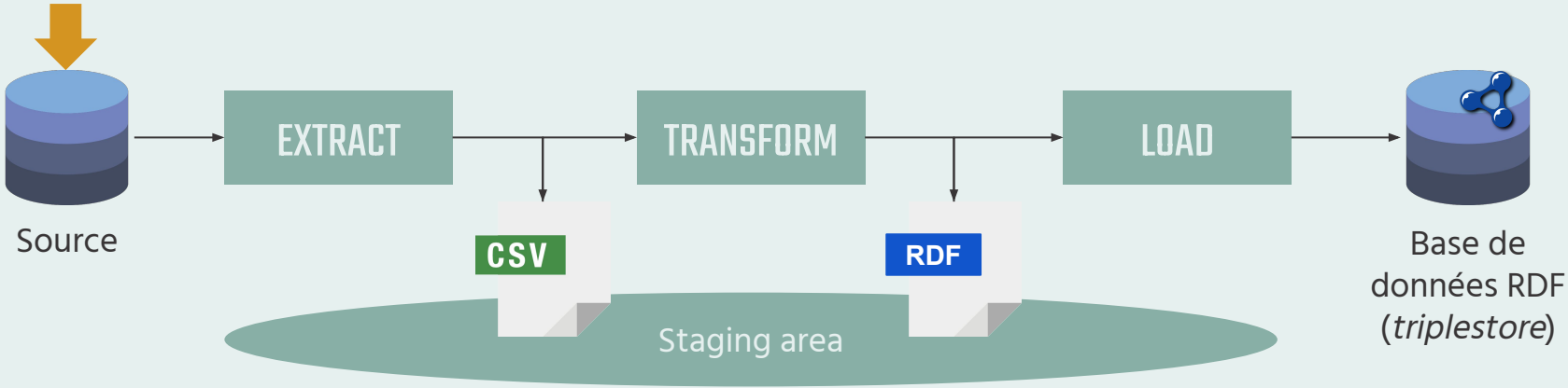
https://nleguillarme.github.io/inteGraph/



Le Guillarme, N., & Thuiller, W. (2023). A practical approach to constructing a knowledge graph for soil ecological research. *European Journal of Soil Biology, 117*, 103497.

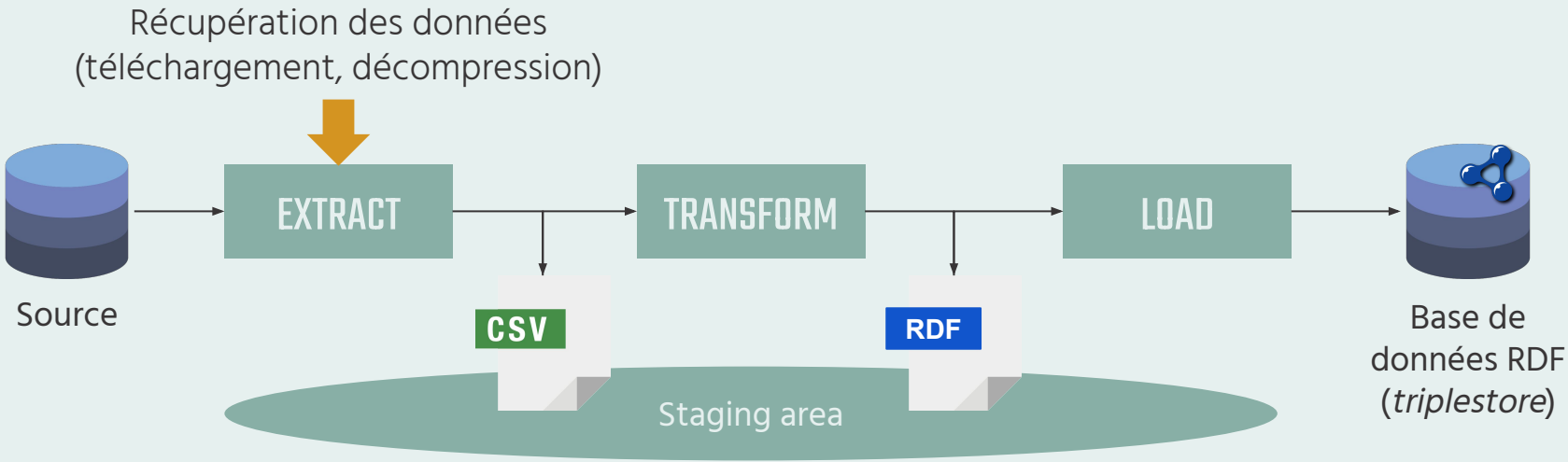
https://nleguillarme.github.io/inteGraph/

- Données structurées
- Fichiers, archives, API
- Stockage local ou source en ligne



Le Guillarme, N., & Thuiller, W. (2023). A practical approach to constructing a knowledge graph for soil ecological research. *European Journal of Soil Biology, 117*, 103497.

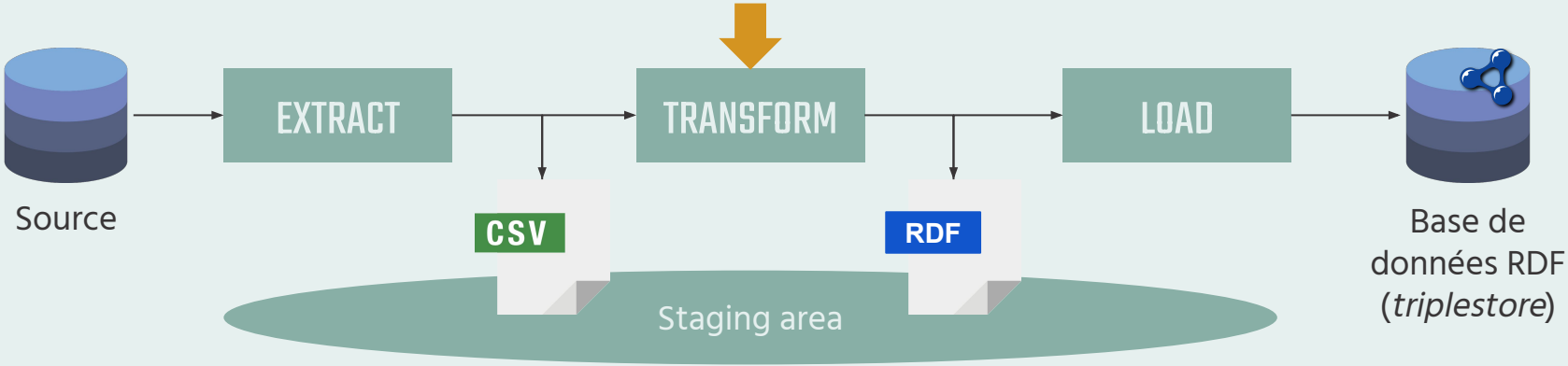
https://nleguillarme.github.io/inteGraph/



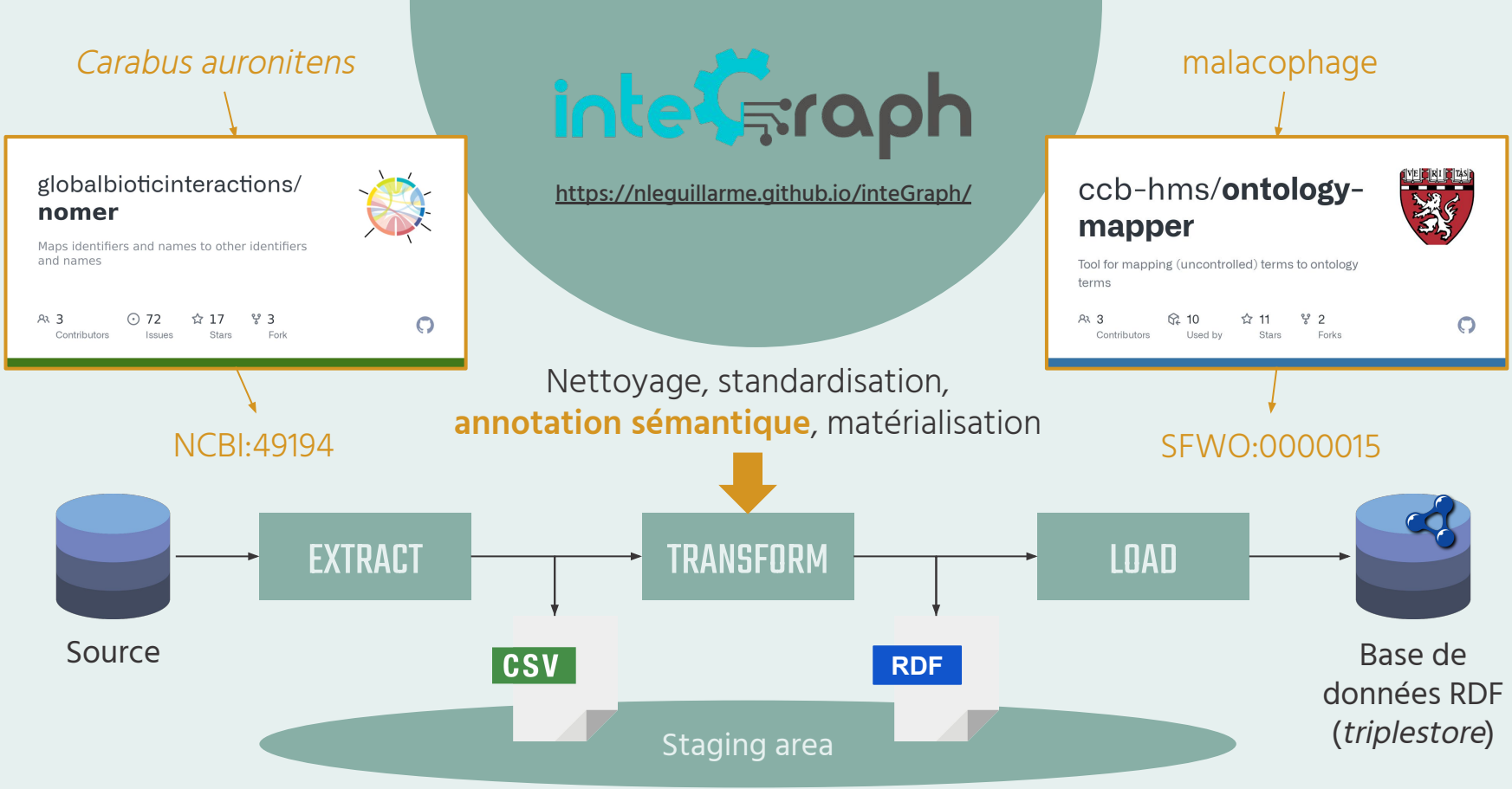
Le Guillarme, N., & Thuiller, W. (2023). A practical approach to constructing a knowledge graph for soil ecological research. *European Journal of Soil Biology, 117*, 103497.

https://nleguillarme.github.io/inteGraph/

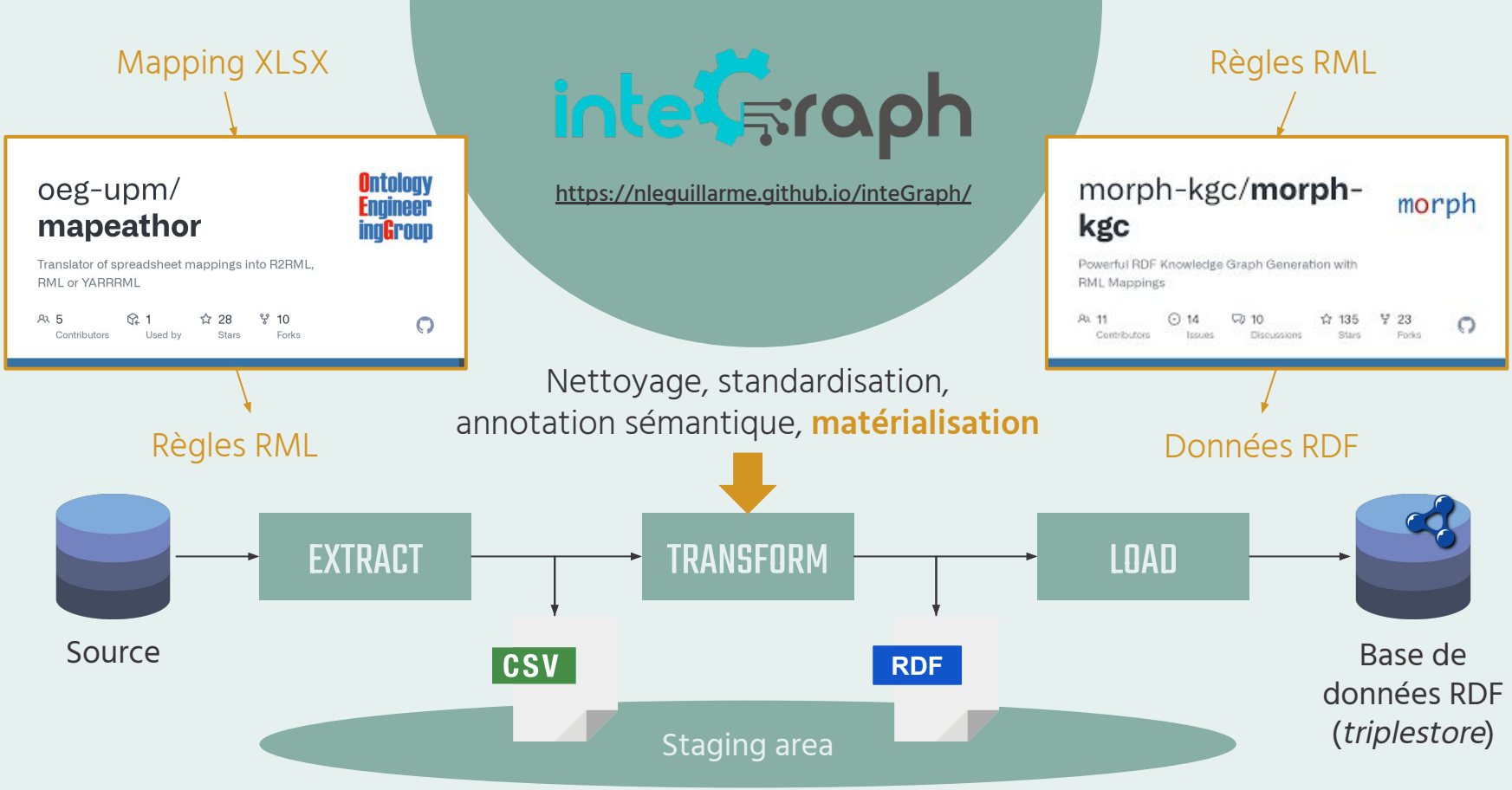
Nettoyage, standardisation,
annotation sémantique, matérialisation



Le Guillarme, N., & Thuiller, W. (2023). A practical approach to constructing a knowledge graph for soil ecological research. *European Journal of Soil Biology*, *117*, 103497.

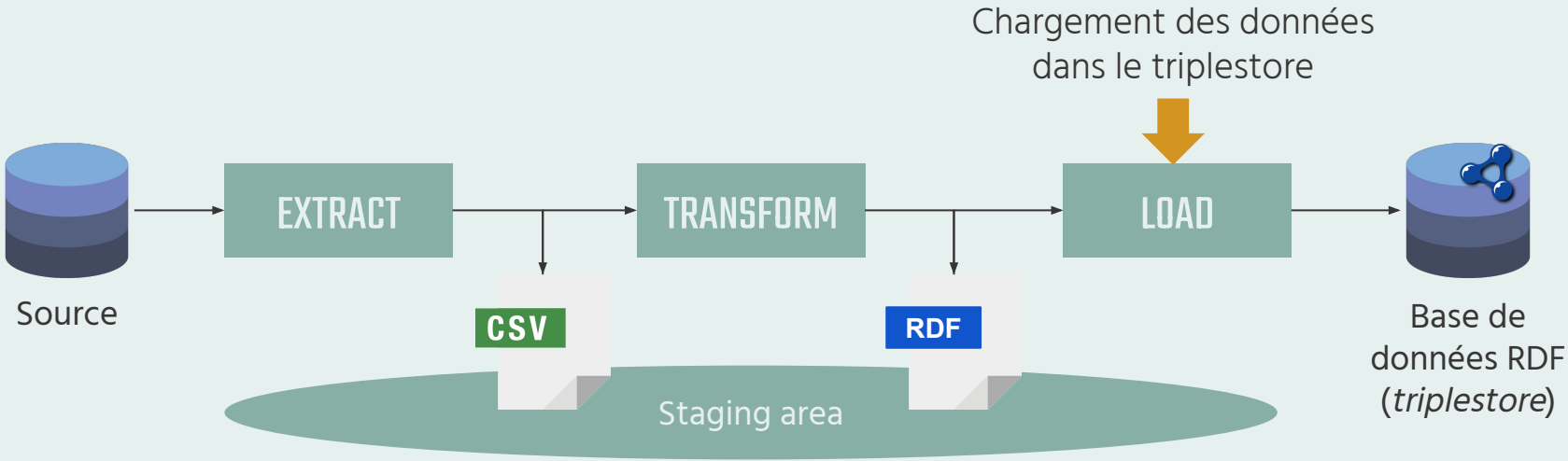


Le Guillarme, N., & Thuiller, W. (2023). A practical approach to constructing a knowledge graph for soil ecological research. *European Journal of Soil Biology, 117*, 103497.

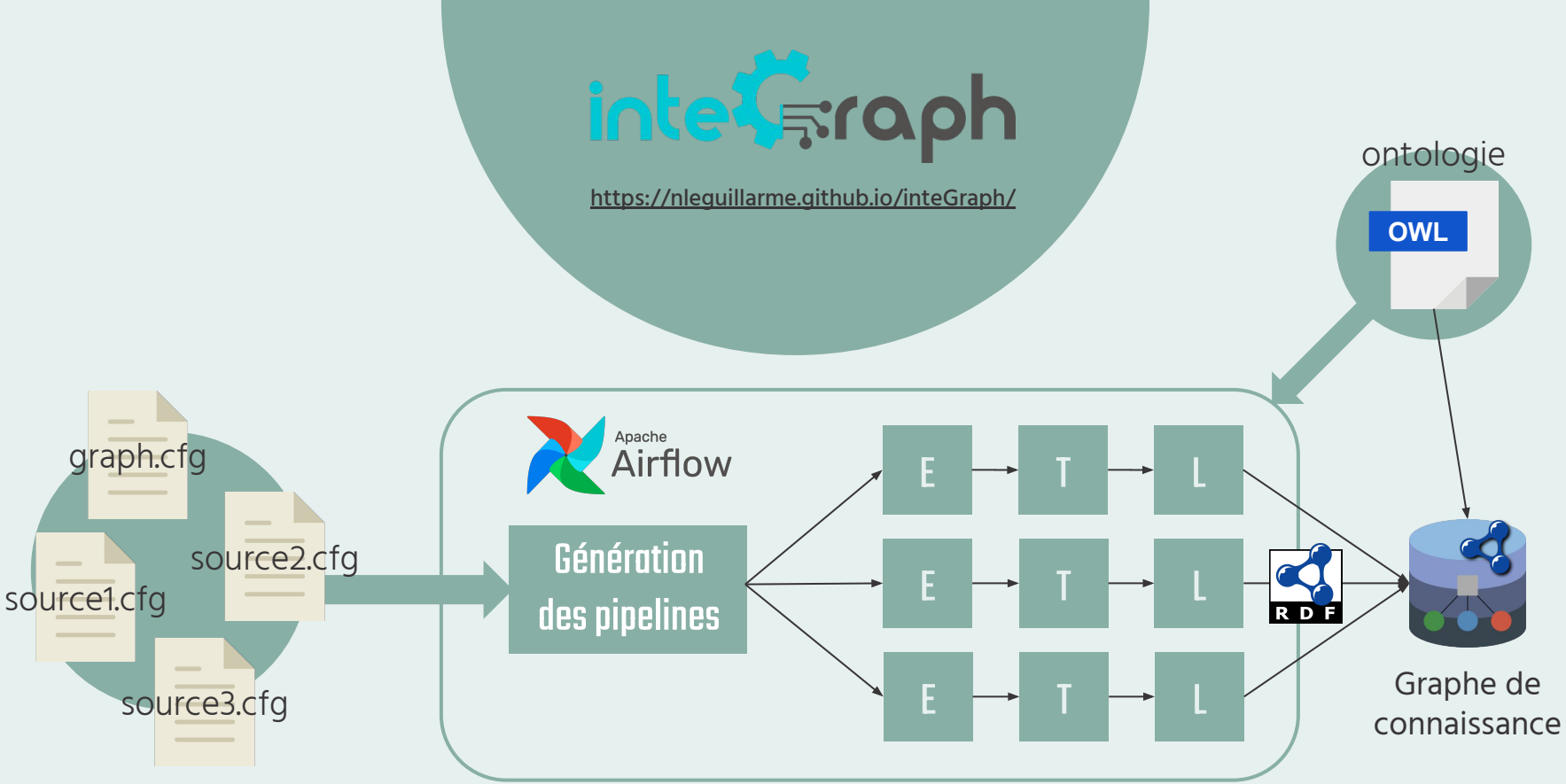


Le Guillarme, N., & Thuiller, W. (2023). A practical approach to constructing a knowledge graph for soil ecological research. *European Journal of Soil Biology, 117*, 103497.

https://nleguillarme.github.io/inteGraph/



Le Guillarme, N., & Thuiller, W. (2023). A practical approach to constructing a knowledge graph for soil ecological research. *European Journal of Soil Biology, 117*, 103497.



Le Guillarme, N., & Thuiller, W. (2023). A practical approach to constructing a knowledge graph for soil ecological research. *European Journal of Soil Biology, 117*, 103497.

GRATIN

A graph of trophic information

SOILFOODWEB ONTOLOGY

NCBI

BETSI

FUNFUN

CSV

GLOBI

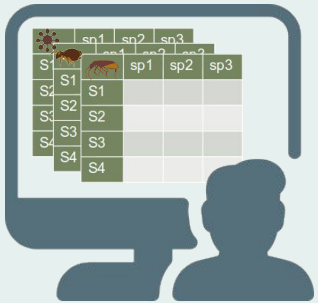
GRATIN est un graphe de connaissances sur l'écologie trophique du sol qui intègre une douzaine de bases de données (bactéries, champignons, protistes, invertébrés, nématodes...)

GRATIN

A graph of trophic information

GratinNavigatoR : un package R pour faciliter l'accès à GRATIN

get.trophic.groups(sciName="Achipteria")



GRATIN est un graphe de connaissances sur l'écologie trophique du sol qui intègre une douzaine de bases de données (bactéries, champignons, protistes, invertébrés, nématodes...)

GRATIN

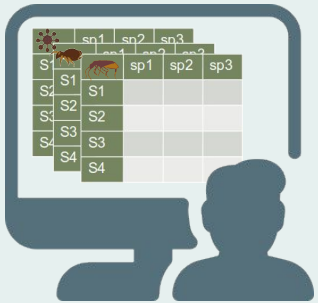
A graph of trophic information

GratinNavigatoR :
un package R pour
faciliter l'accès à GRATIN

get.trophic.groups(sciName="Achipteria")

```
SELECT ?sciName ?groupName WHERE
{
    ?query rdfs:label "Achipteria"
    ?taxon rdfs:subClassOf ?query
    ?organism ro:member_of ?taxon
    ?organism ro:member_of ?groupIRI
    ?taxon rdfs:label ?sciName
    ?groupIRI rdfs:label ?groupName
}
```

SPARQL

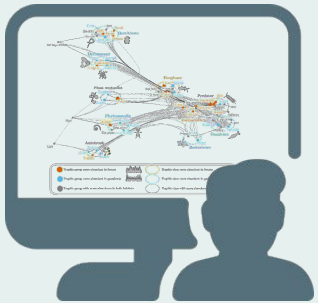


GRATIN est un graphe de connaissances sur l'écologie trophique du sol qui intègre une douzaine de bases de données (bactéries, champignons, protistes, invertébrés, nématodes...)

GRATIN

A graph of trophic information

sciName	groupName
Achipteria coleoptrata	Acari.all
Achipteria coleoptrata	Oribatida.all
Achipteria coleoptrata	Oribatida.detritivores

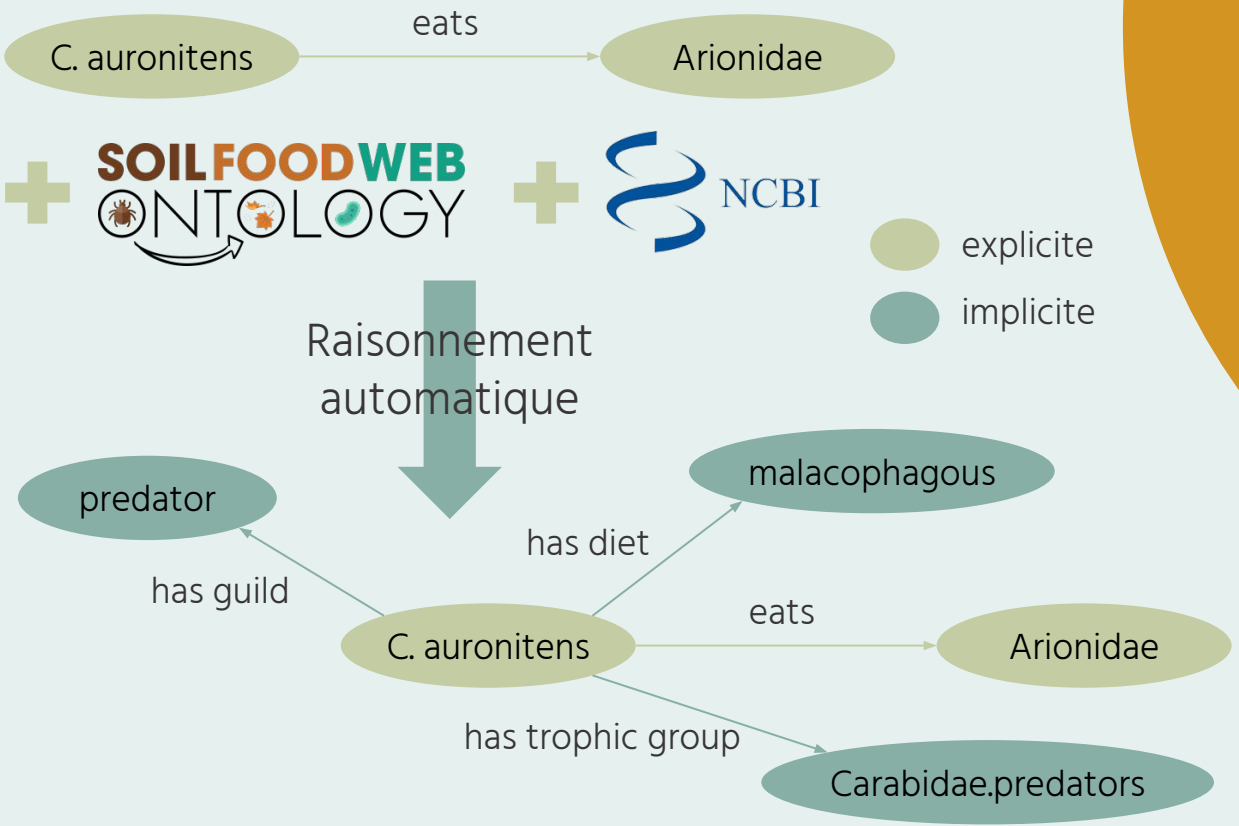


GRATIN est un graphe de connaissances sur l'écologie trophique du sol qui intègre une douzaine de bases de données (bactéries, champignons, protistes, invertébrés, nématodes...)

GRATIN

A graph of trophic information

Inférence dans GRATIN



Inférence dans GRATIN

Il existe plusieurs variantes du langage OWL permettant de trouver un compromis entre l'**expressivité** et l'**efficacité** du raisonnement :

- OWL 2 Full
- OWL 2 EL
- OWL 2 QL
- OWL 2 RL

GraphDB

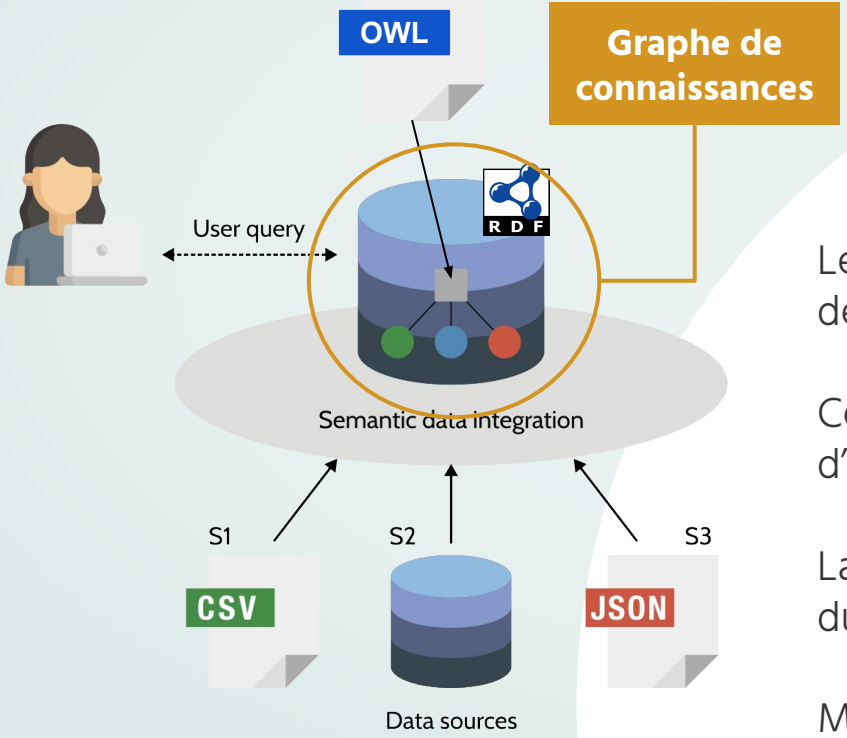
Critères de choix du triplestore :

- Prise en charge du profil OWL 2 RL
- Version gratuite
- Scalabilité

RDFOX
(in-memory)

GRATIN
A graph of trophic information

Pour conclure...



Les graphes de connaissances ont un fort potentiel de facilitation de l'accès à l'information en écologie.

Ce potentiel dépend fortement de la disponibilité d'ontologies spécialisées suffisamment expressives.

La complexité (réelle ou perçue) des technologies du web sémantique rendent leur adoption difficile.

Manque une solution de triplestore libre/gratuite.

MERCI !

Des remarques ? Des questions ?
N'hésitez pas à me contacter par mail :

nicolas.leguillarme@univ-grenoble-alpes.fr

Mickaël Hedde
(Eco&Sols, INRAE)

Wilfried Thuiller
(LECA, CNRS)

Nicolas Le Guillarme
(LECA, UGA)

Irene Calderon Sanou
(LECA, CNRS)