



The LifeWatch ERIC Biodiversity
& Ecosystem eScience Conference

PLOVDIV, BULGARIA

7-10 JULY 2026



KYBELE:

Early progress towards an AI-assisted FAIR
service for biodiversity data integration

Presenter: Stelios Ninidakis, HCMR Greece



08 July 2026



Current Bottleneck

- **Crucial information trapped in full text** (taxonomic treatments/descriptions) and is not machine-actionable.
- **Habitat, trophic guilds and trait information is inconsistent across sources;** taxon names drift across backbones and synonyms.
- **Generic LLMs help extraction but are not reliable for science** without vocabularies, provenance, and evaluation/curation.

Goal: an AI-powered, FAIR-compliant system that extracts and links biodiversity knowledge from literature (molecular and ecological).

Demonstrator: European soil invertebrates (springtails—Collembola) as proof-of-pipeline



4 phases

Data Harvesting

Literature Annotation

Large Language Model

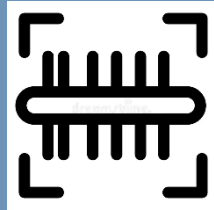
Deploy & Spread

Phase 1: Data harvesting

BOLD SYSTEMS

Manual Curation

DNA barcodes



Metadata

Taxonomy
Traits
Habitats
Trophic guilds
(where available)

1. Find taxonomy synonyms (cross check with different backbones)
2. Analyze literature for alternative writings on traits/TGs/habitats
3. Combine & Cross-check with other backbones

Outputs:

1. A FAIR, domain specific, curated dataset to the community
2. Terminology rules for habitats/TGs/traits



Phase 2: Annotation of BiodiversityPMC

Available at: <https://biodiversitypmc.sibils.org/>

 MEDLINE

 PMC+

 Supplementary data

 PLAZI

 Zenodo

 **40'778'291**
TOTAL DOCS

 **8'478'896**
TOTAL DOCS

 **29'656'484**
TOTAL DOCS

 **1'071'184**
TOTAL DOCS

 **405'000**
TOTAL DOCS

1. Structured **domain-specific vocabularies** built from Phase 1
2. Normalized entities/terms in the corpus
3. Semi-manual curation

+230000 annotations of trophic guilds

Etiquettes de lignes	Medline	Plazi	PMC	PMC-Supp	Zenodo	(vide)	Total général
Carnivores	27517	2786	93334	13736	20716		158097
animal consumers	30	0	123	20	1		174
Carnivores	5764	557	21393	3783	5149		36648
epigeic animal and microorganism consumers	0	0	2	0	0		2
nematode feeders	1	0	4	0	0		5
Predators	21718	2224	71766	9932	15515		121161
zoophages	4	5	46	1	51		107
Detritivores	1679	58	7740	972	668		11117
decomposers	1047	5	5437	495	207		7191
Detritivores	579	32	2025	429	221		3286
Detritus feeders	17	7	126	17	108		275
hemiedaphic microorganism consumers	0	0	1	0	0		1
litter comminutors	0	0	0	0	0		0
litter consumers	2	0	10	1	0		13
saprophages	34	14	141	30	132		351
Herbivores	12717	232	39699	5968	3750		62367
algal feeders	2	0	19	1	16		38
epigeic plant consumers	0	0	1	0	0		1
Herbivores	12208	214	37257	5671	3219		58570
Phycophages	2	0	8	3	0		13
phytophages	51	16	416	75	351		909
plant and microorganism consumers	0	0	2	0	0		2
plant consumers	24	0	107	8	5		144
primary consumers	430	2	1889	210	159		2690
Microbivores	387	14	1485	393	134		2413
bacterivores	191	0	669	180	2		1042
euedaphic microorganism consumers	0	0	2	0	0		2
fungivores	133	11	530	175	83		932
Microbial feeders	3	0	34	1	0		38
Microbivores	32	0	108	8	2		150
mycophages	10	3	69	2	45		129
saprotrophic consumers	0	0	0	0	0		0
secondary decomposers	18	0	73	27	2		120
(vide)							
(vide)							
Total général	42300	3090	142258	21069	25268		233994

Phase 3: AI Component - Current

1. utilize pretrained models (e.g. Llama, Qwen3)
2. use Retrieval-Augmented Generation (RAG) to retrieve accurate information from BiodiversityPMC through APIs
3. perform LLM Question & Answering
4. lightweight fine-tuning (e.g., LoRA)
5. benchmarking-based fine-tuning
6. try to keep low computational footprint



Demonstrator

← → ↺ qa.dev.sibils.org/?q=What+trophic+guilds+are+associated+with+collembola+?&col=pmc&page=1&q 67% ☆

[Hide advanced options](#) | [Show examples](#)

ADVANCED OPTIONS

Mode

Retrieval

Documents to process

Extractive/Generative model + Dense/Sparse search

N-answer synthesis

*Based on the provided documents, **Collembola** are primarily associated with detritivore and herbivore trophic guilds. They play a crucial role in decomposing organic matter and contribute to nutrient cycling within the soil food web. Their functional roles can vary depending on the ecosystem type and land management practices.*

Search results: **What trophic guilds are associated with collembola ?**

MEDLINE 0 **PMC+** 1 PLAZI 0

1 of 1 result

Model: qwen3-8b-fp8

1 Based on the provided documents, Collembola are primarily associated with detritivore and herbivore trophic guilds. They play a significant role in decomposing organic matter and contribute to nutrient cycling within the soil food web . Their functional roles can vary depending on the ecosystem type and land management practices .

Mixed forests with native species mitigate impacts of introduced Douglas fir on soil decomposers (Collembola) Lu et al. Abstract Forest ecosystem management requires the conservation of associated biodiversity. Enriching native forests with economically valuable conifer species provides economic gains and meets the increasing societal demand for timber but may threaten biodiversity. Soil sustains most of forest biodiversity, but the impact of changes in tree species composition, including native and non-native species, on soil invertebrates remains little studied. We investigated the impact of different forest types on the taxonomic and functional composition of springtail communities (Collembola, Insecta), an abundant and diverse microarthropod group inhabiting litter and soil. Using nat

ID: PMC12050398

> See document

Divergence of feeding channels within the soil food web determined by ecosystem type. Understanding trophic linkages within the soil food web (SFW) is hampered by its opacity, diversity, and limited niche adaptation. We need to expand our insight between the feeding guilds of fauna and not just count biodiversity. The soil fauna drive nutrient cycling and play a pivotal, but little understood role within both the carbon (C) and nitrogen (N) cycles that may be ecosystem dependent. Here, we define the structure of the SFW in two habitats (grassland and woodland) on the same soil type and test the hypothesis that land management would alter the SFW in these habitats. To do this, we census the community structure and use stable isotope analysis to establish the pathway of C and N through each

ID: PMC3894882

> See document

Demonstrator is here <https://qa.dev.sibils.org/>

Next plans

Deployment as public service

- **Web portal** for easy access
- **Chatbot** for instant & interactive data discovery
 - Filtering
 - Focusing

Supporting more functionalities

- More structured outputs
- Add new taxonomic groups



Project Values

- **Extensibility:** New taxonomic groups, traits, and scientific domains can be added with limited effort
- **Scalability:** The system can expand to process larger collections of publications, additional taxa, and new types of biodiversity information.
- **Interoperability:** Outputs can be produced using usable, enabling integration with other platforms and research infrastructures.
- **FAIRness:** Data are designed to be FAIR-ready.
- **Traceability:** Extracted information remains linked to its original publication and supporting evidence.
- **Reusability:** Structured outputs can be reused in databases, analytical workflows, knowledge graphs, and AI applications.
- **Reliability:** Controlled vocabularies, validated taxonomic sources, and evaluation procedures support consistent and trustworthy results.
- **Openness:** The project promotes open data, open standards, and broad access to biodiversity knowledge.



Partners



UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II



Italian National
Research Council





The LifeWatch ERIC Biodiversity
& Ecosystem eScience Conference

PLOVDIV, BULGARIA

7-10 JULY 2026

THANK YOU!

Questions?

sninidakis@hcmr.gr

