

Semantics of Biodiversity: from Thesaurus to Linked Open Data (LOD)

Dominique Vachez^{1,2}, Isabelle Gomez^{1,2}

¹[Inist-CNRS](#), France ; ²[GDR SemanDiv-CNRS](#), France

dominique.vachez@inist.fr ; isabelle.gomez@inist.fr

Topics: Semantic interoperability

Keywords: biodiversity, thesaurus, ontology, linked open data, fair terminology, semantic web

[poster abstract first approved for [47th LIBER Annual Conference](#) (04-06 July 2018, Lille, France): Research Libraries as an Open Science Hub]

RESEARCHERS REQUIREMENTS:

Biodiversity data are characterized by their **interdisciplinarity** and syntactic/**semantic heterogeneity**.

They are **disseminated** among many databases without concerted harmonization of their **metadata** for **accessibility** and **reusability**.

It is essential to compensate for the deficit of **controlled vocabularies** and **semantic standards (terminological / ontological)** that are widely recognized in biodiversity and ecology disciplines.

ISSUES:

Achieving the **interoperability** of databases related to scientific research and literature in ecology/biodiversity, using harmonized and **structured** terminological resources (**thesauri, domain ontologies**), enriched with **semantic relationships** to connect concepts and perform logical reasoning on data.

PROJECT SEQUENCE:

Partnership engaged in 2014 between *INIST* and various networking research organizations on the **interoperability** of data in Ecology (*CNRS, INRA, FRB Foundation for Research on Biodiversity*).

- 1st step (2014):

Comprehensive survey of existing **thesauri** in **Environmental Sciences** and their international / European groupings.

Highlighting their complementarity and **alignments** within the **web of data**, in order to: **homogenize** metadata; facilitate indexing and **mining** of heterogeneous and **multidisciplinary data**; make these data interoperable and **shareable (FAIR principles)**.

Findings: lack of thesaurus dedicated to **Biodiversity** subject field and registered in the **LOD cloud**.

- 2nd step (2015):

Development of the *Biodiversity Thesaurus*, **bilingual** and **polyhierarchical**, containing 615 general concepts about ecological components of biodiversity, based on the indexing terms of *CNRS* bibliographic database: grouped in 79 semantic and thematic categories, rich in **synonyms** and lexical variants, **open-access** (*CC BY license*), **interoperable** and **aligned** resource. Compliant with **semantic web** standards (**SKOS/RDF** exchange format, **URI**-concepts) and **ISO international standards** on terminology and thesaurus construction.

With a consistent **hierarchical** structure and **polysemy** resolution, the thesaurus is not only considered as an indexing/search language but as a concept-based **knowledge organization system**.

- 3rd step (2017):

Creation of a **research group** on **biodiversity semantics** involving 40 participants from 18 French and international research structures (*CNRS, INRA, IRD, IRSTEA, MNHN, GBIF...*): [GDR SemaDiv](#), headed by [Eric Garnier](#) (*CEFE/CNRS*).

RESEARCH GROUP OBJECTIVES:

- **Structuring knowledge** in the field of biological diversity to facilitate **semantic annotation** and **data sharing**
- Conceptual basis to stimulate research in **ecoinformatics** and **data integration**
- **Synergy** between actors with complementary approaches: ecologists, information specialists and computer scientists
- Using the *Biodiversity Thesaurus* and *Essential Biodiversity Variables* classes as a conceptual framework for the *GDR*: - formalizing the thesaurus as a lightweight **ontology**; - **mapping** with more specific thesauri (plant/animal traits) and/or domain ontologies; - **interlinking** with research **databases**.

PROJECT EVOLUTION:

- **Displaying** and sharing the *Biodiversity Thesaurus* : - in the platform [Loterre](#) (*Linked Open TERminology REsources*): visualization, browsing and access to a RDF triple store; - in the ontology repository [AgroPortal](#) (*BIODIVTHES*); - in the [FAIRsharing](#) registry
- **Collaborative management** with *GDR* experts: reviewing of **definitions**
- Enrichment by **terminological watch** and **extraction** from specialized corpora
- **Comparative inventory** and **alignments** with other European/international multilingual **semantic resources** (*Agrovoc, GEMET, EnvThes, ENVO...*) acting as hub for interconnections
- Evolving towards an **ontology** of biodiversity
- Aid for **text and data mining**
- Integration into **termino-ontology services** (*EcoPortal*), **ecology databases directories** (*InDoRES/BBEES*), **(meta)data portals** of biodiversity observatories (*French Biodiversity data hub/PNDB*) or environmental **data repositories**.

Bibliography/Webography:

- Dominique Vachez. « [Panorama des thésaurus existants dans le domaine des sciences de l'environnement](#) » et « [Technologies pour la gestion des thésaurus et de leur interopérabilité : standards, outils de gestion et d'alignement](#) » In: *Ecole thématique E-Envir 2019 - Les données ouvertes en sciences environnementales : concepts et méthodes de qualification et de partage des données à l'ère du big data* (28-31 oct. 2019, Gif-sur-Yvette, France)
- Dominique Vachez, Marie-Claude Quidoz. « [T-Semandiv - un nouveau thésaurus en biodiversité](#) » In: *SIST 2018 - Séries Interopérables et Systèmes de Traitement. Qualité des données* (28-29 juin 2018, Guyancourt, France)
- Eric Garnier, Ulrike Stahl, Marie-Angélique Laporte, Jens Kattge, Isabelle Mougenot, et al. "[Towards a thesaurus of plant characteristics: an ecological contribution](#)". *Journal of Ecology*, 2017, 105, pp.298–309
- Dominique Vachez. "[Environmental thesauri: comparisons and alignments. Project for the constitution of a thesaurus of Biodiversity serving research data](#)" In: *IndexMed 2015 Seminar on Methods and tools for the mining of multi-sources and heterogeneous data in ecology* (14 oct. 2015, Marseille, France)
- Marie-Angélique Laporte, Isabelle Mougenot, Eric Garnier, Ulrike Stahl, Lutz Maicher, et al. "[A semantic web faceted search system for facilitating building of biodiversity and ecosystems services](#)". *Data Integration in the Life Sciences - Lecture Notes in Computer Science*, 2014, vol.8574, pp.50-57
- William K. Michener and Matthew B. Jones. "[Ecoinformatics: supporting ecology as a data-intensive science](#)". *Trends in Ecology & Evolution*, 2012, vol.27, no.2, pp.85-93
- [Semantic and Terminology Resources in Environmental and Biodiversity Sciences](#) (Zotero Group Library)