HAL
open science

Comparative study of thesauri in Environmental Sciences - Best practices in thesaurus design and FAIRification

Dominique Vachez

To cite this version:

Dominique Vachez. Comparative study of thesauri in Environmental Sciences - Best practices in thesaurus design and FAIRification. 2021. hal-03264850

HAL Id: hal-03264850

https://hal.science/hal-03264850

Preprint submitted on 18 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Comparative study of thesauri in Environmental Sciences
Best practices in thesaurus design and FAIRification

Dominique Vachez
dominique.vachez@inist.fr
INIST-CNRS, 2 rue Jean Zay CS 10310 F-54519 Vandoeuvre-les-Nancy, France

[Translated from: "*Etude comparative de thésaurus en Sciences de l'Environnement - Bonnes pratiques de conception et FAIRisation de thésaurus, D. Vachez*"]

Abstract: We conducted a detailed comparative analysis of the structure of several thesauri in the field of environmental sciences (EnvThes, GEMET, OZCAR-Theia) with the aim of highlighting their limitations and complementarities.
The criteria (both qualitative and quantitative) selected to compare these terminological resources are listed and evaluated so as to propose a guideline of best practices for the design and utilization of thesauri as open semantic repositories and thus ensure the interoperability of environmental research data.
In order to promote the sharing and reuse of (meta)data, the vocabularies chosen to represent the domain knowledge must indeed conform to a set of principles and rules developed at the international level (ISO standards, FAIR principles, web of data recommendations, exchange formats).
In addition to this study, we also made a comparison with six other thesauri in the topics of agronomy, ecology and biodiversity (Biodiversity Thesaurus, AGROVOC, GACS, EARTh, AnaEE-France and INRAE Thesaurus).

Keywords: thesaurus, environmental sciences, terminology resources, semantic interoperability, knowledge organization system, best practices, Linked Open Data, FAIR vocabulary

A] Comparative analysis of three thesauri in the field of the Environment

We compared three freely available environmental science thesauri (**EnvThes, GEMET, OZCAR-Theia**) according to different criteria related to their semantic quality and their compliance with:

- **Linked Open Data** (Web of data) recommendations (W3C, 2006/2010)

- **SKOS RDF** data model for Simple Knowledge Organization Systems (W3C, 2009)

- **ISO 25964** standards for thesauri and their interoperability (2011/2013)

- **ISO 704** standard on principles and methods of terminology work (2009)

- **FAIR Principles** (2016)

As a reminder, ISO 25964 defines a thesaurus as a controlled and structured vocabulary, with a combination of hierarchical and associative relationships between **concepts**, as well as intra/inter-linguistic equivalence relationships between the **terms** representing these concepts in the same language (preferred term and synonyms) or other languages (translations).

SKOS is a standard format that is part of Semantic Web technologies, and whose model is based on a minimal set of axioms and integrity conditions. It can be applied to various simple knowledge organization systems, such as thesauri.

To represent each **concept** in a given natural language, one preferred **term** (skos:prefLabel) is chosen as a designation (ISO 704), also known as a descriptor (ISO 25964), which is therefore standardized but comes as close as possible to the specialized language used in a subject field.

This preferred label may be complemented by one or more synonyms/quasi-synonyms (alternative labels skos:altLabel), also called equivalent or non-preferred terms.

When correspondence relations (matching) are created between concepts belonging to different thesauri, we speak of alignments or mappings between semantic resources (skos:mappingRelation). These matches (or conceptual equivalences) can be accurate (skos:exactMatch) or approximate (skos:closeMatch/relatedMatch). Indeed, in thesauri, definitions are not established as formally as in ontologies.

Nevertheless, since the publication of these standards, and if they are properly followed, a thesaurus is no longer seen merely as an indexing language for information retrieval, but becomes a real concept system that can be considered as a lightweight ontology in knowledge engineering.
With a few transformations based on an extension of the SKOS model according to the ISO 25964 standard (iso-thes), it is also possible to re-engineer thesauri into simple ontologies.

From this comparison, we will deduce a certain number of best/good (or less good) practices in the construction of thesauri, helping us to guide the choice of reference terminologies intended to enrich metadata in the environmental field, and to ensure a semantic coherence that fosters the integration and reuse of data.

Terminological clarification: we see here that the term equivalence can be used in various senses
→ equivalence between *terms* of one vocabulary
(intra-linguistic: **synonymy** / inter-linguistic: **translation**)
→ equivalence between *concepts*: **alignment** (equivalence mapping between different vocabularies)

Main criteria for thesaurus comparison

- Terminological richness (terms):

- number of **preferred terms** (skos:prefLabel)
- number of **synonyms** (**equivalence** relationships)
  (**alternative** terms-labels / skos:altLabel / **USE** tag / **UF** tag: used for)
- number of spelling **variants** (**hidden** terms-labels / skos:hiddenLabel)
- specialized language

- Semantic relations (concepts): **hierarchical** and/or **associative** relationships

- number of **broader** concepts (skos:broader / **BT** tag: broader term)
- number of **narrower** concepts (skos:narrower / **NT** tag: narrower term)
- number of **top** concepts (**TT** tag: top term)
- number of hierarchical levels (hierarchical **depth**)
- number of sibling concepts (under the same broader)
- number of **related** concepts (skos:related / **RT** tag: related term)

- **Hierarchies**: consistency; mono/**polyhierarchy**; 3 types (**generic** / **partitive** / **instantial**)

- **Grouping of concepts**: collections, microthesauri

- **Linguistic equivalences** between terms - **Multilingualism**

- Presence of **definitions** (multilingual)

- French, European or **international** reputation

- **Uniform Resource Identifier** (URI) for each concept

- Links with the **web of data** (LOD - **Linked Open Data** cloud)

- **Alignments** made with other thesauri (**equivalence** mapping relations)

    - **Incoming** and/or **outgoing** links (**reciprocity**)
    - Display the nature of these links (skos:**exact** / close / broad / narrow or related**Match**)

- Free **searchable**, **downloadable** and **interoperable** resource (**SKOS-RDF** format)

- Exploration / navigation (ergonomics) / query **interfaces** / **SPARQL** endpoint / API

- **License** of (re)use (**free/open**)

- Maintenance and **versioning** of the resource

Note: In the following text, expressions in **Latin bold type** correspond to **qualities**, while expressions in ***italics*** correspond to ***limitations***.

1] EnvThes (Environmental Thesaurus)

European thesaurus developed within the **LTER-Europe** network (**Long-Term Ecosystem Research in Europe**) (**EnvEurope / ExpeER** project) since 2014

- generalist and related to the **INSPIRE** Directive's themes

- based on ontologies (OGC O&M, OBOE and SERONTO), as well as on the **US LTER Controlled Vocabulary** thesaurus and the environmental keywords of Microsoft Academic Research

- used by DEIMS-SDR (**Dynamic Ecological Information Management System** - Site and Dataset Registry) for annotation and quering metadata purposes, and implemented as a semantic backbone for data sets within the European project ECOPOTENTIAL

- recommended for the geocatalog of RZA metadata (French network of **Zones Ateliers** member of **eLTER / ILTER**) and the OSU OREME data portal (Environment Research Observatory of Montpellier)

- thesaurus base URI = http://vocabs.lter-europe.net/EnvThes/10000

To be noted: This thesaurus has evolved significantly since an initial comparative study in 2014 and contains many deprecated concepts.

EnvThes Qualities / Advantages

- semantic richness: **2805 non-deprecated concepts** (gathered under 7 super-concepts)
- **multilingualism** (26 languages) under development (English predominant)
- 1690 **alternative terms** (altLabel) English/French

- presence of multilingual **definitions**
- **polyhierarchical** structure
- **TopBraid**: main navigation interface with open and **multilingual** access (***preferred terms***)
- **EcoPortal**: interface displaying only **validated concepts**; access by ***English*** **synonyms**
- **URI** concepts (http://vocabs.lter-europe.net/EnvThes/21921): **opaque** and **dereferenceable** identifiers
- **SKOS-RDF** format
- complete thesaurus downloadable in RDF ttl (GitHub, EcoPortal)
- **numerous** outgoing **alignments** to AGROVOC, GEMET, EARTh, US-LTER, EuroVoc, DBpedia
- SPARQL **endpoint** (CEH)
- present in the **FAIRsharing** catalog

EnvThes Limitations / Disadvantages

- *no grouping* in *collections*
- ***mix*** of ***deprecated*** (2930) and validated (***difficulty for alignments***) concepts sometimes put in relation *skos:related*
- ***deprecated*** concepts are systematically linked ***skos:exactMatch*** with the *replacing concepts*, but also with concepts from *other thesauri*
- *mismanagement of* ***synonyms*** → ***duplicate*** concepts
  [e.g. "*rainforest*" / "*rain forest*"]
- *mismanagement of* ***translations*** → ***synonyms without a prefLabel*** in the same language (e.g. "*endangered species*" or "*vascular plant species*" have ***no translation*** of the *prefLabel* but present ***altLabels*** in many languages)
- *misuse* of property ***skos:altLabel*** frequently used instead of ***translations*** (see above) or for *morpho-syntactic* ***variants*** *(instead of skos:hiddenLabel, e.g. "pine forested"; "environmentally polluted")*
- ***spelling mistakes****: e.g. skos:prefLabel "permittivtiy"@en*
- ***few French*** *terms* (only 208 French preferred labels)
- ***EcoPortal*** *interface*: - ***unilingual query*** (English), *not searchable in French;*
  - ***translations without language tags****;*
  - *query* ***bugs****:* ***uniterms*** or ***multiterms*** *sometimes* ***not found*** *(e.g. "biodiversity", "biomass", "genetic diversity");* ***incomplete*** *answers and/or not following the* ***alphabetical*** *order;*
  - ***bugs*** *on the hierarchical tree display (e.g. "object of interest");*
  - some ***synonymous*** *terms remain* ***identical to the preferred term*** *(e.g. "arbre" (fr) /trees);*
  - ***non-display of narrower concepts*** *in the Details*
- ***TopBraid*** *interface*: - *not searchable by* ***synonyms****;*
  - *hierarchical display* ***mixing*** *English and French terms* (when the term is *multilingual*)*;*
  - ***incomplete answers*** *(search only performed in* ***Starts with*** *mode)*
- ***EcoPortal/TopBraid***: *none of them allows to* ***retrieve all*** *the terms* ***containing*** a character string (e.g. **diversit**) [***BARTOC*** Skosmos browser *being* ***out of date*** *(**2017**)*];
  - when ***displaying a concept, the hierarchical tree does not unfold completely****:*
  therefore it is ***not possible to visualize*** *an entire* ***polyhierarchy*** *(e.g. event "algal bloom"; parameter "plant nutrient")*
- ***SKOS-RDF anomalies***: *thesaurus* of both *skos:Concept* and skos:ConceptScheme types;
  - ***no skos:inScheme*** property at Concept level;
  - *misuse of alignments (**xMatch**)* between concepts *within the same thesaurus*
- ***no SKOS-RDF export*** *at* ***individual concept level***

- presence of ***html code*** *into* ***text notes*** of the *rdf file*
- many ***empty*** *definitions*
- ***extra whitespaces*** in alternative labels
- ***non-reciprocal*** *alignments (**no incoming alignments,** except Biodiversity and OZCAR thesauri)*
- ***no*** usage ***license***

The SKOS Testing Tool (**qSKOS**) can detect a rather large number of anomalies or inconsistencies with the **SKOS specification** (**SKOS Reference** W3C recommendation) and/or the **SKOS Primer** user guide.

Some errors present in 2018 have been fixed in the latest GitHub release, such as *synonyms that were identical to prefLabel* (which nonetheless still appear in *EcoPortal*).

- ***concepts without any label:*** e.g. *http://vocabs.lter-europe.net/EnvThes/22021*
- ***orphan concepts*** without any semantic relation with other concepts:
  e.g. *http://vocabs.lter-europe.net/EnvThes/30020*

Most of these "concepts" often cumulate several flaws (*orphan concept*, *no preferred label*, ***non-opaque URI***): e.g. *http://vocabs.lter-europe.net/EnvThes/water_pressure*

- ***omitted language tags*** (labels or definitions)
- ***homonymous labels:*** identical ***prefLabel*** in the same language for distinct concepts
  e.g. "*eau douce*" (fr); "*évaporation*" (fr); "*demography*" (en)
- concepts ***related both hierarchically*** *and* ***associatively*** (e.g. "*aquatic fungi*" BT/RT "*fungi*")
- ***lack*** *of* ***inverse*** *relationships* (broader/narrower)
- ***lack*** *of* ***symmetric*** *relationships* (related)
- ***associative*** *relations between sibling concepts of the* ***same*** *level*

2] GEMET (GEneral Multilingual Environmental Thesaurus)

European thesaurus developed by the **European Environment Agency / European Environment Information and Observation Network (EEA / Eionet)** since 1997

- generalist and compatible with all the themes of the **INSPIRE** directive

- resulting from the compilation of several semantic resources (lexicons, thesauri)

- commonly recommended for metadata enrichment: OSU OREME, BBEES/InDoRES (Bases de données Biodiversité, Ecologie, Environnements, Sociétés)

- thesaurus base URI = http://www.eionet.europa.eu/gemet/gemetThesaurus

GEMET Qualities / Advantages

- covers all environmental topics
- semantic richness: **5530 concepts**
- 4 super-groups + 32 groups + 40 themes (collections)
- Eionet interface (multilingual): access to **top-concepts** allocated to each group
- AgroPortal interface (in English): navigation in the **hierarchical tree of concepts** from the **110 top-concepts**
- LusTRE **(Linked Thesaurus fRamework for Environment)** interface: **multilingual federated search** among several thesauri (GEMET, AGROVOC, EARTh, EuroVoc)
- TemaTres interface: **alphabetical** list and **hierarchical** tree
- **polyhierarchical** structure

- **multilingualism** (37 languages)
- presence of **definitions**
- **reciprocal alignments** with AGROVOC, EuroVoc
- **SKOS-RDF** format
- **URI** concepts (http://www.eionet.europa.eu/gemet/concept/892): **opaque** and **dereferenced** identifiers
- complete thesaurus **downloadable** in SKOS RDF (Eionet Portal / AgroPortal)
- open **license CC BY**
- SPARQL endpoint (EEA)
- **version** history
- present in **FAIRsharing** catalog

GEMET Limitations / Disadvantages

- ***no SKOS-RDF export at individual concept level***
- ***rare English synonyms*** *(alt* or *hiddenLabel);* ***no French synonym***
- ***identical French*** *prefLabels* for ***different concepts*** (distinct definitions and English labels): e.g. "*aménagement du territoire*" (4 concepts); "*pollution du sol*" (2 concepts in NT/BT)
- ***identical prefLabel*** *for a concept and its* ***broader***: "*aménagement du territoire*" (French); "*tutkimus*" (Finnish) …
- some cases of ***unilingual*** *descriptors:* "*silviculture*" *(en)*
- ***confusion*** between the labels of certain concepts, groups and themes (e.g. "*Biosphere*")
- semantic ***incoherence*** at the same hierarchical level: e.g. BT "*biosphere*" /NT "*anatomy; biological process; ecology; evolution; organism; virus*"
- ***hierarchical redundancies***: e.g. "*watercourse*", "*sea*", "*water reservoir*" are narrower concepts of both "*hydrosphere*" and "*water (geographic)*" (while the latter is itself narrower than "*hydrosphere*")
- non specialized (average *hierarchical depth*: 5 levels)
- *no inverse relationships* ***skos:topConceptOf***
- ***Eionet*** *interface****: no browsing possible*** *in a* ***global hierarchical tree***
- ***AgroPortal*** *interface*: ***- no navigation possible by groups*** *or* ***themes;***
  - unilingual query (English) not searchable in French;
  - when displaying a concept, the hierarchical tree does not unfold completely: therefore does not allow to identify a polyhierarchy *(e.g. "watercourse")*
- ***LusTRE*** *interface*: ***- no browsable hierarchical tree; - search*** *only performed in* ***Starts with*** *mode****; - outdated*** *version (****2012****)*
- ***TemaTres*** *interface*: - ***internal URIs not corresponding to thesaurus basic URIs***;
  - unilingual query (English); - *no groupings*; - ***not updated*** *since* ***2012***

The SKOS quality check tool (**qSKOS**) can detect several other anomalies:

- ***cyclic hierarchical relationship***: a concept is both broader and narrower of the same concept. "*ecosystem assessment*" is ***both BT and NT*** of "*ecosystem boundary*"
- ***orphan concepts***: "*biological diversity*"
- ***homonymous preferred labels***: e.g. "*waste water treatment plant*" (English) and other concepts in many other languages
- ***collisions*** *between* ***hierarchical*** *and* ***associative semantic relations***
- ***associative*** *relations between sibling concepts of the* ***same level***

3] OZCAR-Theia thesaurus (Critical Zone Observatories - Application and Research)

French thesaurus developed by the information system of the **Theia** data cluster (IR **Data Terra** / **Earth System** – **Continental surfaces**) and the **OZCAR** infrastructure (**eLTER-RI**) since 2019

- specialized on a range of environmental variables

- based on the **GCMD Earth Science** Keywords vocabulary (Global Change Master Directory, NASA) and upon certain groups of parameters of the French Sandre framework on water chemistry reference data

- thesaurus base URI = https://w3id.org/ozcar-theia/ozcarTheiaThesaurus

OZCAR-Theia Qualities / Advantages

- semantic richness: **430 concepts** gathered under 2 super-concepts + 2 collections
- **numerous alignments** (exact, close or relatedMatch) with **EnvThes, AGROVOC, GACS, GCMD, EARTh, GEMET, AnaEE thesauri …**
- **Skosmos** consultation interface **(browsing** by **hierarchical tree** and by **groups)**
- individually exportable **concepts** in **SKOS-RDF** format
- **URI** concepts (https://**w3id.org**/ozcar-theia/biosphere): **permanent** and **dereferenceable** identifiers
- resource **downloadable** directly in rdf (or by query of the SPARQL endpoint)

OZCAR-Theia Limitations / Disadvantages

- ***monolingual*** (English): *French absent*
- possible confusion between collection titles and top terms
- lack of some global concepts (*critical zone*; *geosphere*)
- complex terms sometimes far from natural language or encompassing *several concepts* e.g. "***COHV - Solvents - Freons (groundwater)***" *(hindering the alignments)*
- many ***homonymous*** terms (differentiated by brackets)
- ***heterogeneous*** spellings
- ***incorrect*** spellings
- some *identical terms* appear several times (with the same concept URI) in the hierarchical tree, but refer to ***different concepts***: e.g. "*precipitation amount*"
- ***absence*** *of* ***associative relationships*** *skos:related*
- ***absence*** of ***synonyms***
- ***absence*** of ***definitions***
- ***no*** usage ***license***
- URI concepts: ***meaningful*** identifiers (~prefLabel) → risk for long term sustainability
- ***metadata*** not filled in (resource, concepts); no creation/modification ***dates***
- ***alignments***: some ***exactMatch*** are actually *closeMatch* or *relatedMatch*.
- alignments with ***deprecated*** **EnvThes** or **EARTh** ***concepts***
- ***not present*** in ***FAIRsharing***

The quality of SKOS has been tested with the Loterre (**L**inked **o**pen **ter**minology **re**sources) web service, as well as with the SKOS Play! qSKOS online service, by unchecking the "Broken links" box (to avoid a proxy error, considering the numerous alignments). These services do not always detect the same type of failings and can therefore be usefully complementary.

It should be noted that many of these anomalies were undetectable or untraceable in Skosmos.

- URI appearing as a member of the *Variable categories* collection, but ***non-existent as a concept*** (no skos:Concept, no label) and without any hierarchical relationship, thus ***invisible in Skosmos***
- URI appearing in rdf:Description, but only by its alignments, without skos:Concept, nor skos:prefLabel, nor skos:inScheme, thus ***invisible in Skosmos***
- URI appearing as skos:narrower , but missing as skos:Concept. It ***cannot be found*** in the alphabetical list and appears as "***null"*** in the ***Skosmos*** *hierarchy*
- ***reflexive hierarchical relationship***: the concept is its *own broader*
- ***missing inverse relations*** (unidirectional broader/narrower): in general, the absence of one of the 2 inverse relations (skos:broader/skos:***narrower***) in the rdf file, causes display failures in the unfolding of the Skosmos hierarchical tree (e.g. "*Precipitation amount*" does not appear under "*Liquid precipitation*")
- ***homonymous labels*** appearing several times in the hierarchy

➔ After consultation with the Theia/OZCAR information system team, most of these defects could be corrected in a new version of the thesaurus.

Note: In addition to the checking tools provided by **qSKOS** and **Loterre**, if the thesaurus has been loaded into the **VocBench** open source editor, it is possible to use an integrated module for testing the validity of SKOS **integrity** constraints (ICV - Integrity Constraint Validator).

B] Good / Best practices (structure and description of thesaurus)

▪ Thesaurus (class skos:ConceptScheme)

- **Concept scheme**: **semantic resource**, semantic **artefact** or **knowledge organization system** (KOS).

▪ Concepts (class skos:Concept)

- Indicate the membership of each concept (= **conceptual resource**) to the thesaurus (= concept scheme) with the property **skos:inScheme**.

▪ Hierarchies (properties skos:broader/narrower)

- **Homogeneity** in hierarchical relationships between concepts: avoid confusion between hierarchy (skos:broader/narrower) and grouping (skos:Collection / skos:member).

- The relationships skos:broader and skos:narrower are **inverse** of each other and it is better to have them both in the rdf file.

- Hierarchical **coherence**: specific concepts of the same hierarchical branch should not belong to fundamentally different **semantic categories** (property, discipline, organism, substance, place, process, action, method...).

- These fundamental categories should be treated either as higher-level general concepts (top-concepts) or as collections of concepts.

- According to **ISO 25964** and **704** standards, we can distinguish 3 types of hierarchies:

- **generic** (is a / sort of / kind of / type of) = *subsumption* relationship
- **partitive** (part of / component of / constituent of / located in)
- **instantial** (*individual* concept / instance of / named entity)

These relations can be expressed using the **subproperties** of skos:broader /narrower from **iso-thes** ontology, respectively:

- isothes:**broader /narrowerGeneric** (**BTG** / **NTG** tags)
- isothes:**broader /narrowerPartitive** (**BTP** / **NTP** tags)
- isothes:**broader /narrowerInstantial** (**BTI** / **NTI** tags)

- **Polyhierarchy** is allowed (mixing various hierarchy types).

Note: Hierarchical relationships of different typologies can be interlinked but they are not necessarily transitive. The skos:**broader/narrower** properties that establish a *direct* hierarchical link between two concepts are therefore not declared transitive in the SKOS model.
If required, it is nevertheless possible to express the *transitivity* of a hierarchical relationship (*direct or indirect*) with the properties skos:**broaderTransitive** / skos:**narrowerTransitive**.

- Limit the number of **top level** concepts to facilitate navigation in the thesaurus.

- Properties linking a thesaurus to all its top-concepts (**skos:hasTopConcept**) must be included in the rdf file inside skos:ConceptScheme; the corresponding inverse properties (**skos:topConceptOf**) appear inside each top skos:Concept.

▪ Associations (property skos:related)

- The associative relationship is **symmetrical** and must therefore be present in both directions.

- This property is **not transitive**.

- The **SKOS** specification does not allow 2 concepts to be linked *associatively*, if they are already linked *hierarchically* (**disjoint** properties).

- Two concepts can generally be associated if they have **similar** or **overlapping** meanings ("*coastal zone*" RT "*littoral zone*") or if the first concept participates in the definition of the second, without being in a hierarchical relationship ("*ecosystem*" RT "*ecosystem services*"; "*soils*" RT "*pedology*").

- **Sibling concepts**: *Sibling* concepts that appear at the same level under the same broader concept in a *whole-part hierarchy* should not be systematically associated with each other. Too frequent associations in the same hierarchy could also indicate hierarchical inhomogeneity or confusion between hierarchies and grouping of concepts.

- An associative linkage can be an opportunity to qualify the nature of this relation and thus to move towards an **ontologization** of the thesaurus.
Some thesauri (e.g. AGROVOC) have subdivided the **skos:related** property by creating several **subproperties** intended to **refine** the associative relationship into various **ontological relations**: cause/effect, agent/action, parameter/process, object/property, technology/product, etc.

▪ **<u>Collections</u> (**class **skos:Collection)**

- **Groupings** of concepts by categories, themes or sub-domains will be managed as collections.

- Each collection consists of a **non-hierarchical** list of concepts (**skos:member**). It is however possible to nest collections and sub-collections.

- Groups can also be managed using the **subclasses** of skos:Collection from the **iso-thes** ontology (isothes:**ConceptGroup** and/or isothes:**ThesaurusArray**) and structuring them into subcollections (subgroups) with the properties isothes:superGroup / isothes:subGroup.

▪ **<u>Definitions</u> (**property **skos:definition)**

- The definition of the concept is optional but strongly recommended to prevent any ambiguity.

- **Unambiguous** definition : if a concept has several definitions, this may mean that we are dealing with different concepts.

- When writing a definition, certain rules have to be observed (**ISO 704**: Terminology work - Principles and methods). *Circular, imprecise* or *negative* definitions must be avoided.
Thus, a concept cannot be defined by means of a second concept that is itself defined using the term designating the former concept. It is also not permitted to repeat the designation of the concept to introduce the definition.

- The definition to be preferred is the **intensional definition**: it consists in stating a more generic (*superordinate*) concept and identifying a combination of delimiting characteristics that distinguish the concept to be defined from other specific (*subordinate*) concepts.

- Give the **source**(s) of the definition.

- Assign a **language code** to each definition.

- SKOS allows to choose a non-literal value as the object of the definition (e.g. the URI of a resource or the URL of a document, instead of a string).

- It is also possible to use the **skos:scopeNote** property in order to clarify and limit the meaning or use of a polysemous term for a given concept in the thesaurus domain: "note that defines or clarifies the semantic boundaries of a concept".

<u>Notes</u>: In its model, the ISO 25964 standard specifies that **ScopeNote** class is associated (*definesScopeOf*) with **ThesaurusConcept**, while **Definition** is associated (*isDefinitionOf*) with **ThesaurusTerm**. As a result, the content of Definition could go beyond the restricted meaning attributed to the term within the thesaurus, and could also authorize to add a definition to a non-preferred term.
Nevertheless, the W3C SKOS model (2009) has not been updated after the release of the ISO standard (2011) and both the skos:scopeNote and skos:definition properties can be used to annotate a skos:Concept.

▪ **<u>Preferred terms</u> (**property **skos:prefLabel)**

- No more than **one preferred lexical label** (term) for each concept and in each language (**SKOS Reference** constraint) [normative]

- Do not designate ambiguously different concepts with the same preferred term (**SKOS Primer** recommendation). Indeed, **homographic** labels are conflicting, whether they are homonymous or polysemous.

- Avoid choosing one preferred term that concatenates or covers **several** concepts at once, whether they are separated by punctuation marks (hyphens, commas) or conjunctions (and, or).
In fact, unlike a classification, the *precoordination* of multiple notions in one same *compound* term (multiterm) should remain exceptional in a thesaurus (unless it is ultra-specialized and dedicated to a restricted user community).

- Assign a **language code** to each term.

- The lexical labeling should be **precise** enough but not too complex in order to facilitate **alignments**.

- The use of **parentheses** should be limited and is mainly used to disambiguate or qualify the preceding term when it is polysemous or is homonymous with another term. When possible, it is preferable to use a more precise term: e.g. "*trophic chain*" (ecology) vs. "*food chain*" (food industry).

- Multi-word terms should be built in **natural language order**. An excessively long label that is far from natural language may be more like a definition or represent a *combination* of several concepts.

- The preferred label of a concept can be changed in a new version (use the **skos:changeNote** property) while keeping its URI.

- The preferred terms of a vocabulary must not vary from one terminology repository to another (always showing the latest version).

▪ **<u>Language equivalences</u>**

- Promote **multilingualism** (at least bilingualism) with an equivalent term in each language for each preferred term, unless **exact cross-language equivalence** does not exist for every language.

- In some situations (named entities, Latin names, loan terms), identical labels may be used in different languages.

▪ **<u>Synonyms</u> /<u>alternative terms</u> (**property **skos:altLabel) or <u>hidden terms</u> (skos:hiddenLabel)**

- It is recommended to enrich the thesaurus with equivalent or synonymous terms (facilitating **querying** and/or **alignments**):

  - Prefer exact synonymy (strict equivalence) to approximate synonymy (or quasi-synonymy)
  - **Quasi-synonyms** are allowed even if they are only interchangeable in certain contexts
  - **Acronyms** are most often placed as **synonyms**, unless they are widely used and the terminology is sufficiently specialized to avoid ambiguity (DNA in biology, PCB in chemistry...)
  - Possibility of **inverting** synonyms and preferred terms (term depreciation, evolution of uses, disambiguation): to be specified in a historical note (**skos:historyNote**)
  - Detect identical concepts (*duplicates*) that could be **deprecated** and provide new synonyms.

- The ISO 25964 standard allows a more *specific* term to be placed in an *equivalence* relationship by designating it with a more *generic* term, in order to avoid multiplying the number of preferred terms, while retaining an entry term. However, it should be considered that an alternative term (skos:altLabel) that is more precise than the preferred term (skos:prefLabel) could represent a more specific **emerging** concept (skos:narrower).

- Query results can be improved by the presence of multiple **hidden variant forms** or **misspellings** that are searchable but not displayed (**skos:hiddenLabel**).

Note: The **SKOS eXtension for Labels** (SKOS-XL) allows to treat labels (lexical entities) not simply as literal strings, but as RDF resources (**skosxl:Label**) and to link them together through **skosxl:labelRelation** properties.
This makes it possible to create sub-properties of skosxl:labelRelation such as the **acronym** of an expanded form.

▪ Concept deprecation

- Not all thesauri deal with concepts in the same way when these become obsolete.

These rejected concepts can be deleted entirely or reported in a new version of the thesaurus using annotations on the concepts that replaced them (synonymy *inversions*, concept *splitting* or *merging*).

- Several potential situations:

  - Total **deletion** of the deprecated concept, with possible mention of its past existence in the **annotation** of another concept (**skos:historyNote** or **skos:changeNote**);
  - **Maintaining** the deprecated concept and its URI, but with a preferred label absent or containing "*(deprecated)*" or "*(obsolete)*" in order to clearly distinguish it and exclude alignments in exactMatch;
  - AGROVOC: keeping the URI of the deprecated concept, with an annotation skos:changeNote but without any label;
  - EnvThes: keeping the URI and preferred label of the deprecated concept, under the generic concept "*deprecated concept*" and exactMatch/closeMatch mapping to the URI of the new concept(s).

- If the deprecated labels remain in the SKOS file, the exposure tools (Skosmos, EcoPortal...) can be set to display them or not in the alphabetical/hierarchical visualization interface.

▪ Alignments (property skos:mappingRelation)

- Avoid using of alignment relations (skos:xMatch) to link concepts belonging to the *same* thesaurus.

- Among the different relations skos:exact / close / broad / narrow or relatedMatch , only the equivalence in **skos:exactMatch** is a **transitive** property. It can thus be used to link exactly concepts from several vocabularies in an indirect way, making sure that this equivalence is not approached.

- (Semi) automatic alignments are done on a morpho-syntactic basis, involving **distance/similarity** algorithms between preferred or non-preferred terms. A rereading is therefore essential if the alignment is made through (quasi) synonyms.

- In particular, the treatment of **uniterms** with different meanings from one domain to another (that only broader concepts can disambiguate) may be questionable:
e.g. concept "*antagonism*": **GEMET** **BT** "*chemical property*" / **AGROVOC** **BT** "*biological competition*" are ***wrongfully*** put in *exactMatch*, while they are in fact distinct concepts.

- When a thesaurus refers to a very **specialized** domain and does not contain sufficiently global top concepts to match automatically, it is still possible in a **bottom-up** approach to manually perform **broadMatch** mappings with concepts of a more general-oriented pivot resource.

▪ Identifiers (concept URI)

- **Persistent** identifier starting with a permanent/stable domain name, and ending with a non-explicit (ideally **random** and/or **opaque**) character sequence for each concept.

- The URI of a concept should normally start with the **base URI** of the thesaurus.

- It is important not to modify these URIs over time. Concept URIs should not vary between versions or exposition platforms. It is possible to **deprecate** them, but URIs are not reusable for new concepts.

- **HTTP URIs** must be **dereferenceable**. If the site is **secured**, it will also be necessary to ensure that http URIs are correctly redirected to **https** URLs.

- *Warning*: Some platforms such as TemaTres (searchable but not maintained) impose their *own URI identifiers* to navigate in resources. These URIs should not be used for alignments.

▪ Formats

- Choice of **SKOS (RDF)** format for data exchange (W3C **standard**).

- Possibility of downloading (**concept** level and **resource** level) in various formats and serializations: SKOS (RDF/XML, Turtle...), JSON-LD, CSV, PDF...

▪ Thesaurus metadata

- Metadata about terminology resource are elements to be placed inside skos:ConceptScheme tag.

- Possibility to download separately the **metadata** describing the resource (e.g. AgroPortal).

- Minimum indications (**Dublin Core /** DCMI Metadata Terms):
  - description of the resource (dc:description / rdfs:comment)
  - name of the resource (dc:title / rdfs:label)
  - subject of the resource (dc:subject)
  - version (owl:versionInfo)
  - creator (dc:creator) or contributor (dc:contributor)
  - date of creation (dct:created) and modification (dct:modified) (resource + concept level)
  - uniform + persistent identifiers at resource level (http URI + DOI dc:identifier).

- Other RDF vocabularies applicable to metadata and facilitating interoperability: **DCAT** (Data Catalog Vocabulary) and/or in conjunction with **VoID** (Vocabulary of Interlinked Datasets) more specifically adapted to linked datasets in RDF, and **PROV-O** for provenance informations.

- Choice of a **free user license**:
e.g. Creative Commons type: CC 0 or CC BY or CC BY-SA (cc:license)
with an attribution organization, responsible for the resource (cc:attributionName).

▪ **Web indexing / SEO**

- Improvement of the crawling of the web site hosting the resource for search engines.

- Optimization of **response times** (a thesaurus/KOS federated search tool such as BARTOC with a short *timeout* may disqualify some terminology service API).

▪ **Exposition**

Make sure that all terminology repositories/registries provide access to the **latest version** of the thesaurus.

We have endeavored to apply this set of good practices by constituting the **Biodiversity Thesaurus** at Inist-Cnrs (not included in this study, but appearing in the comparative table in the Appendix), made available on Loterre (exposing via **Skosmos** browser and API) and AgroPortal (BIODIVTHES), and registered in FAIRsharing.

With particular focus on:

-- **bilingualism** (exhaustive)

-- richness of **synonyms** and **hidden terms**

-- groupings into several types of **collections**

-- **polyhierarchy** (according to ISO 25964)

-- **associations**

-- **persistent identifiers** (**ARK** type URIs)

-- **alignments** in the Linked Open Data cloud

-- **ergonomics** of the user interface

-- **FAIRness** (data and metadata): interoperability (**SKOS**) and reusability (**CC BY** license)

-- multiple **downloading** formats: RDF/XML, Turtle, JSON-LD, CSV, PDF

In the Appendix, a **comparative table** illustrates the semantic richness and reusability of 8 open access controlled vocabularies (Biodiversity Thesaurus, AGROVOC, GACS, EnvThes, GEMET, EARTh, AnaEE-France and INRAE Thesaurus) on the concept of "**biodiversity"**, following 24 criteria.

This table is the 2021 update of a comparative survey carried out in 2014 which resulted in the decision to develop a new bilingual thesaurus gathering the general key concepts relevant to the ecological component of biological diversity, within the framework of a Research Group on Biodiversity Semantics (**GDR SemanDiv**):
Semantics of Biodiversity: from Thesaurus to Linked Open Data (LOD)

To find examples of detailed comparisons of the management of semantic relations concerning other **environmental concepts** and/or **thesauri**, refer also to:
Overview of existing thesauri in the field of environmental sciences

Conclusions:

The strengths and weaknesses highlighted for each thesaurus indicate that these terminology resources should be able to **complement** each other, leveraging their specialization, multilingualism and synonymies.

Their accessibility, openness and interoperability in the **web of data** make them suitable **standard** semantic tools, in a context of **open science** and **FAIR principles** development.

The choice of international **pivotal** thesauri**,** sufficiently generalist and rich in exactMatch **alignments** (such as **EnvThes** or **AGROVOC**), can also make it possible to establish **semantic hubs**, with transitive **interconnections** between several **ontoterminological** resources in the interdisciplinary Environment domain.

In addition, these vocabularies will allow the production of terminology-based **semantic annotations** for scientific abstracts or research datasets, exploiting synonymies, alignments or hierarchical expansion of encountered terms.

Finally, they will be used to link environmental publication portals and (meta)data portals through shared **indexings**.

Sources

Best Practice Recipes for Publishing RDF Vocabularies (2008)
https://www.w3.org/TR/swbp-vocab-pub/

SKOS Simple Knowledge Organization System Reference (2009)
https://www.w3.org/TR/skos-reference/

SKOS Simple Knowledge Organization System Primer (2009)
https://www.w3.org/TR/skos-primer/

Linked data (Tim Berners-Lee, 2009)
https://www.w3.org/DesignIssues/LinkedData.html

Key choices in the design of Simple Knowledge Organization System (SKOS) (2013)
https://arxiv.org/pdf/1302.1224

Best Practices for Publishing Linked Data (2014)
https://www.w3.org/TR/ld-bp/

Towards the reuse of standardized thesauri into ontologies (WOP, 2014)
https://www.researchgate.net/publication/272794487_Towards_the_reuse_of_standardized_thesauri_into_ontologies

Turning FAIR into reality (European Commission Expert Group on FAIR Data, 2018)
https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_1.pdf

Technologies for thesaurus management and interoperability: standards, management and alignment tools (2019)
https://e-envir.sciencesconf.org/data/pages/J2_CM4_TechnologieThesaurus_VachezDominique.pdf

SEMANTIC AND TERMINOLOGICAL RESOURCES IN ENVIRONMENTAL SCIENCES (2020)
https://hal.archives-ouvertes.fr/hal-02907484v2/file/RESSOURCES SEMANTIQUES ET TERMINOLOGIQUES EN SCIENCES DE L'ENVIRONNEMENT_09.10.2020.pdf

Semantics of Biodiversity: from Thesaurus to Linked Open Data (LOD) (2020)
https://hal.archives-ouvertes.fr/hal-02907484v2/file/Semantics_of_Biodiversity_from_Thesaurus_to_LOD_VachezDominique_GomezIsabelle_2020.pdf

D2.5 FAIR Semantics Recommendations - Second Iteration (Hugo, Le Franc et al., FAIRsFAIR 2020)
https://doi.org/10.5281/zenodo.4314321

Ten Simple Rules for making a vocabulary FAIR (2020)
https://arxiv.org/pdf/2012.02325.pdf

Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web (Garijo et al.,2020)
https://arxiv.org/pdf/2003.13084.pdf

A Semantic Web methodological framework to evaluate the support of integrity in thesaurus tools
https://doi.org/10.1177%2F0165551519837195 (2020)

APPENDIX

▪ The four main rules of the Web of Linked Data

(Linked Data, *Tim Berners-Lee,* 2006/2009)

1) **identify** resources with **URIs**

2) use **http** URIs (**dereferenceable**) that can be used to **access** information about the resources

3) when dereferencing a URI, return **structured** data respecting **semantic web standards: RDF** and **SPARQL**

4) **link** URIs together to create a **network of** links and discover new information

+ **Open / free licenses** (e.g. CC BY) → **Linked Open Data (LOD)**

▪ Linked Open Data: 5 Stars (2010)

https://5stardata.info/en/

Semantic resources will be able to obtain from 1 to 5 stars in the Linked Open Data (LOD) grading scale as defined by T. Berners-Lee.

* Data **freely** accessible on the **web in any** format (e.g. PDF, JPG), with mention of an **open data license**;
** Data in a **structured**, machine-readable format (e.g. XLS);
*** Open **non-proprietary** format (CSV, XML, ODS ...);
**** **URIs** to **identify** each resource and **W3C open standards** (RDF* languages) to represent them (RDFS, OWL, SKOS, JSON-LD ...) or query them (SPARQL);
***** Data linked to other RDF data via **alignments** between their URIs (**LOD cloud**)

▪ Five Stars of Linked Data Vocabulary Use (2014)

http://www.semantic-web-journal.net/content/five-stars-linked-data-vocabulary-use

Zero: Linked data but no mention of the vocabulary used
* **Dereferenceable information** describing the vocabulary
** **Axiomatization of** vocabulary (**W3C** languages: RDF*, OWL ...)
*** **Outgoing links** to other vocabularies (**alignments**, equivalentClass)
**** **Dereferenceable metadata** on vocabulary (OMV, VOAF...)
***** **Incoming links** from other vocabularies

▪ **FAIR Principles** (2016)

The 4 core **FAIR principles** (Findability, Accessibility, Interoperability, Reusability) share characteristics with the 4 Linked Data basic rules: **persistent identifiers**, **standard access protocols**, **formats** and **languages interoperable** with other resources.

FAIR data must be properly described with **persistent, standardized metadata** and released with a clear data **usage license.**

To meet semantic **interoperability** principle, FAIR (meta)data must use **FAIRified terminologies** / controlled vocabularies for knowledge representation.

FAIR principles have their own ontology: FAIR Vocabulary

Comparative table of 8 thesauri about Environment and Biodiversity (concept "biodiversity")

CONCEPT BIODIVERSITY Thesaurus	# Narrower concepts	# Broader concepts	Level in hierarchy	# Synonyms altLabel (en / fr)	# Hidden terms hiddenLabel (en / fr)	# Sibling concepts (same BT)	# Associated concepts (related)	Multilingual	Definitions	# Groups of concepts (Collection)	Concept URI (LOD dereferencia-tion)	Mapping exactMatch	Incoming links (reciprocity)	Platform	Linked Open Data (SKOS/RDF format)	Skosmos browser	SPARQL endpoint	LOD Cloud	Download at resource level	Download at concept level	License	FAIRsharing	Origin	# Terms (pref+syno) containing *diversit*
Biodiversity Thesaurus	**6**	**2**	2 / 4	**11**	**30**	**29**	**7**	**YES (2) English / French**	1	**3**	http://data.loterre.fr/ark:/67375/BLH-FHNG3BCR-H	3 (agrovoc; gemet; envthes)	NO	AgroPortal	**RDF/XML Turtle JSON-LD**	https://www.loterre.fr/skosmos/BLH/	https://www.loterre.fr/sparklis/		https://www.loterre.fr/science-de-la-vie-sante/biodiversite-2/	RDF/XML Turtle JSON-LD	**CC BY**	https://doi.org/10.25504/FAIRsharing.A29ckB	French	**24**
AGROVOC	5	1	4 / 5	6	0	4	2	**YES (32)**	1	0	http://aims.fao.org/aos/agrovoc/c_33949	7 (**earth**; **gemet;** nalt; **eurovoc**)	**earth**; **gemet;** **eurovoc**	AgroPortal	**RDF/XML NT**	http://www.fao.org/agrovoc/fr/search	https://agrovoc.uniroma2.it/sparql	https://lod-cloud.net/dataset/agrovoc	http://www.fao.org/agrovoc/releases	RDF/XML Turtle JSON-LD	**CC BY SA**	https://doi.org/10.25504/FAIRsharing.anpj91	international	9
GACS	3	1	5 / 6	3	0	**29**	**8**	**YES (22)**	1	1	http://id.agrisemantics.org/gacs/C1918	3 (**nalt**; agrovoc; cabt)	**nalt**	AgroPortal	**RDF/XML Turtle**	http://browser.agrisemantics.org/gacs/en		https://www.lod-cloud.net/dataset/GACS	https://github.com/gacs/gacs-scheme/raw/master/src/gacs-core-scheme.ttl	RDF/XML Turtle JSON-LD	**CC BY**	https://doi.org/10.25504/FAIRsharing.p1dodf	international	9
EnvThes	**6**	1	6 / 7	0	0	**31**	1	**YES (24)**	1	0	http://vocabs.lter-europe.net/EnvThes/21673	6 (agrovoc, earth; eurovoc; gemet; lter)	NO	EcoPortal	**Turtle**	https://bartoc-skosmos.unibas.ch/envthes/en/	http://vocabs.ceh.ac.uk/edg/tbl/sparql		https://github.com/LTER-Europe/EnvThes/tree/master/CurrentVersion	NO	unknown	https://doi.org/10.25504/FAIRsharing.dS2o69	European	**30**
GEMET	4	1	6 / 8	0	0	22	2	**YES (37)**	2	2	http://www.eionet.europa.eu/gemet/concept/827	2 (**agrovoc**; eurovoc)	**agrovoc**; earth; envthes; eurovoc	AgroPortal	**RDF/XML**	NON	https://semantic.eea.europa.eu/sparql	https://lod-cloud.net/dataset/gemet	https://www.eionet.europa.eu/gemet/en/exports/rdf/latest	NO	**CC BY**	https://fairsharing.org/bsg-s001444/	European	11
EARTh	2	1	5 / 6	1	0	8	6	**YES (2) English / Italian**	1	2	http://linkeddata.ge.imati.cnr.it/resource/EARTh/27980	3 (**agrovoc**; eurovoc; gemet)	**agrovoc**; envthes	TemaTres	**RDF/XML**	https://skosmos.dev.finto.fi/earth/en/	http://linkeddata.ge.imati.cnr.it:8890/sparql	https://lod-cloud.net/dataset/environmental-applications-reference-thesaurus	https://old.datahub.io/dataset/environmental-applications-reference-thesaurus	RDF/XML	CC BY NC ND	NO	European	12
AnaEE-France Thesaurus	**12**	1	2 / 6	0	0	7	0	**YES (2) English / (French)**	2	0	http://opendata.inra.fr/anaeeThes/c3_2393	2 (agrovoc; gemet)	NO	AgroPortal	**RDF/XML**	https://bartoc-skosmos.unibas.ch/anaeethes/en/	http://opendata.inra.fr:8080/openrdf-sesame/repositories/anaeeThes		http://data.agroportal.lirmm.fr/ontologies/ANAEETHES/submissions/6/download	RDF/XML Turtle	**CC BY**	https://doi.org/10.25504/FAIRsharing.49bmk	French	11
INRAE Thesaurus	**9**	0	1 / 2	0	0	top concept	0	**YES (2) French / English**	0	2	http://opendata.inrae.fr/thesaurusINRAE/c_5836	1 (agrovoc)	NO		**RDF/XML**	https://consultation.vocabulaires-ouverts.inrae.fr/thesaurus-inrae/fr/			https://consultation.vocabulaires-ouverts.inrae.fr/rest/v1/thesaurus-inrae/data	RDF/XML Turtle JSON-LD	**Open Licence**		French	**36**