Séminaire SIST 2018

Thesauform

un outil collaboratif pour faciliter la création de vocabulaire contrôlé par des experts de domaine

Marie-Claude Quidoz (CNRS/CEFE)

GDR SEMANDIV

FRB FOUNDATION FOR RESEARCH ON BIODIVERSITY

CESAB CENTRE FOR THE SYNTHESIS AND ANALYSIS OF BIODIVERSITY

CENTRE D'ECOLOGIE FONCTIONNELLE & EVOLUTIVE

Quelques définitions

- Un vocabulaire contrôlé est une liste de termes (mots et expressions) soigneusement choisis pour désigner les concepts d'un domaine (un seul terme préférentiel et éventuellement plusieurs entrées non-préférentielles). Il réduit l'ambiguïté inhérente au langage humain naturel, où différents noms peuvent être attribués à un même concept

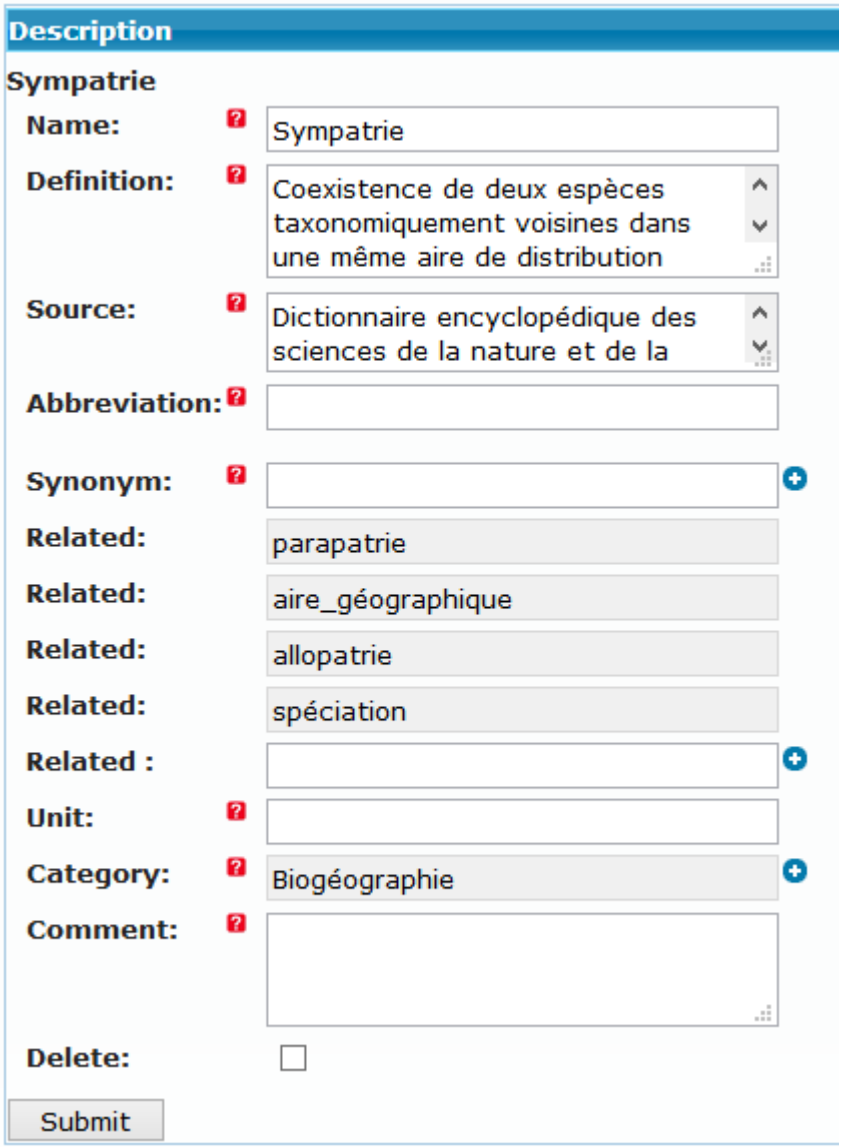
- Un thésaurus permet d'organiser et de structurer un vocabulaire contrôlé à partir de relations sémantiques entre concept (de types hiérarchique ou associatif) et d'équivalence entre termes

Thésauform, c'est quoi

- Un outil simple pour répondre à une question simple mais qui s'appuie sur des technologies du web sémantique

- Un outil collaboratif pour faciliter la création / gestion de thésaurus
  - ✔ Élaboration collaborative des termes qui seront soumis ensuite pour annotation (ajout et/ou modification) aux experts de confiance
  - ✔ Procédure de vote mise à la disposition des experts de confiance pour validation des propositions suite à des propositions préétablies.

Atout 1 de Thesauform : la simplicité



- Le langage est simple, adapté au chercheur (et pas au spécialiste de l'ingénierie des thésaurus)
- Les intitulés SKOS sont remplacés par des expressions « parlantes » pour les chercheurs
  - label = name ; TA = related ; TG = category
- Des rubriques sont exprimées différemment pour plus de clarté
  - description est scindée en deux (définition et source)

Atout 2 de Thesauform : le vote

Vote for the term: Liste_rouge_uicn_2018

Back to voting home page
your tally for this term: 0

Vote for the options using the drop-down list. Each of your votes will be recorded in the tally score above.
Please click the + icon if you want to leave a comment. When you have finished, click OK.

How important is this term for inclusion? no response Comment:

Please rate the following proposals linked to the term :

Name

Current name: Liste_rouge_UICN_2018 no response Comment:

Definition

Current definition: Liste rouge de l'UICN en 2018 (ref: wiikipedia) no response Comment:

Proposition 2: Liste rouge de l'UICN en 2018 (ref: www.wikipedia.fr) no response Comment:

Abbreviation

Current abbreviation: UICN2018 no response Comment:

Category

Current category: Liste_rouge_uicn no response Comment:

Quelques autres atouts

- Logiciel OpenSource
  - ✔ CC BY-SA 4.0
  - ✔ https://github.com/CESAB-FRB/Thesauform
- Différentes possibilités d'export
  - ✔ Si authentifié
    - ► en XLS (liste des termes ainsi que la hiérarchie)
    - ► en SKOS étendu (interopérable)
  - ✔ Sinon présence de nombreuses API

Parmi les inconvénients

- Utilisé par une communauté de faible taille
- Prototype en Java développé dans le cadre d'une thèse en écoinformatique (2008-2011)
- Il ne respecte pas en totalité la norme ISO 25964 (2011 & 2013)
  - ✔ Import possible uniquement en CSV
  - ✔ Les URI ne sont pas déréférençables
  - ✔ Impossibilité de gérer la polyhiérarchie
  - ✔ Absence de multilinguisme
  - ✔ Absence de facette

Des exemples d'utilisation

- BETSI
  - Traits fonctionnels des invertébrés du sol
  - Contrôle de cohérence des données saisies dans la base de données BETSI avec le thésaurus T-SITA
- GDR SemanDiv
  - Evaluer les 80 définitions du thésaurus en faisant appel à des experts du domaine

THE CESAB THESAUFORM

A COLLABORATIVE TOOL FOR THE DOMAIN SCIENTIST

Alison Specht (CESAB-FRB), Baptiste Laporte (CESAB-FRB),
Marie-Angelique Laporte (Biodiversity International), Eric Garnier (CEFE-CNRS)

Common and understandable terminology is essential for sharing data and information. The lack of clear and accessible terminology and definitions of those terms is very evident in a synthesis centre, where our business is to assist researchers to aggregate existing data. In some fields, such as in the identification and definition of plant traits, we are well progressed, but in others much less so, and in order to advance the world of open data it is increasingly urgent that we improve the semantic capture of domain-specific data.

This web-based tool allows a group of scientists to collaborate in the identification of the main concepts used in their domain, to propose definitions for the terms related to the concepts, and establish a list of synonyms for these terms.

The CESAB Thesauform, developed from a previous version (Laporte* et al., 2012) allows :

- Concept categorisation,
- Enrichment : add / delete terms,
- Modification of concepts (terms and their associated information),
- Coordination of different points of view : voting protocol, and
- Graphic information display.

IMPLEMENTATION

The Thesauform is designed to allow domain specialists to define the terminological aspects of specialised domain concepts. The definition of the key concepts is accomplished collaboratively.

These collaboratively-defined concepts are added to a SKOS thesaurus with metadata information, conforming to the Dublin Core Abstract Model and structured according to property and value (respectively P and V in Figure 2).

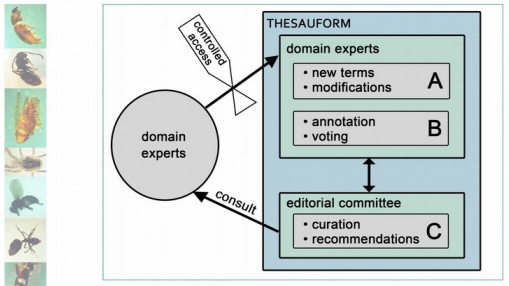


Figure 1: Workflow between domain experts and the Thesaurus. Letters refer to stages in the process illustrated accordingly

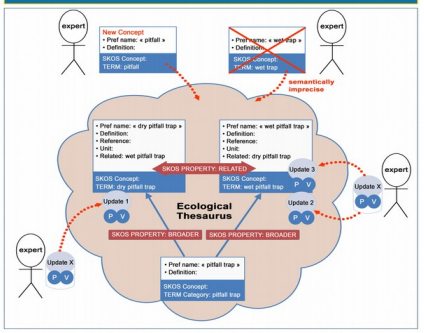
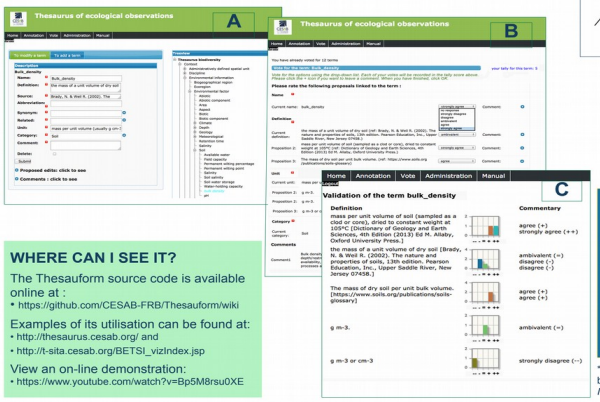


Figure 2: Thesauform structure. P = Property, V = Value



a selection of macro-invertebrates caught in wet pitfall traps

WHERE CAN I SEE IT?

The Thesauform source code is available online at :

- https://github.com/CESAB-FRB/Thesauform/wiki

Examples of its utilisation can be found at:

- http://thesaurus.cesab.org/ and
- http://t-sita.cesab.org/BETSI_vizIndex.jsp

View an on-line demonstration:

- https://www.youtube.com/watch?v=Bp5M8rsu0XE

NEXT STEPS

The THESAUFORM will be used to:

- deliver concepts (terms, definitions, sources, related terms etc) to ontologies (e.g.ENVO, PCO) to be validated by the community.
- provide a framework for use by different domains.
- publish the resulting thesauri on-line.

* Laporte M-A, Mougenot I., Garnier E. (2012) Thesauform–Traits: a web based collaborative tool to develop a thesaurus for plant functional diversity research. *Ecological Informatics* 11: 34-44.

Contact : cesab@fondationbiodiversite.fr
Web site : www.cesab.org and www.fondationbiodiversite.fr

FRB
FONDATION POUR LA RECHERCHE SUR LA BIODIVERSITÉ

Si vous avez des questions

Rendez-vous devant le poster