Collaborative management of Semantic Web datasets: maintaining EuroSciVoc with VocBench and ShowVoc

Anikó Gerencsér[1,*], Baya Remaoun[1] and Enrico Bignotti[2]

[1]Publications Office of the European Union, 20 rue de Reims, 2417, Luxembourg
[2]Infeurope SA, 113 Rue Adolphe Fischer, 1521, Luxembourg

Abstract

This paper presents the collaborative maintenance approach for the European Science Vocabulary (EuroSciVoc) through the implementation of VocBench and ShowVoc corporate knowledge management systems. EuroSciVoc is a standardized taxonomy designed to classify research projects within the CORDIS platform of the Publications Office of the European Union. This paper demonstrates how VocBench and ShowVoc, two open-source platforms leveraging Semantic Web technologies, enable efficient management and dissemination of controlled vocabularies through their specialized features and collaborative functionalities.

Keywords

SKOS, Taxonomies, Alignments, Linked Open Data, Knowledge management

1. Introduction

As (controlled) vocabularies provide a shared understanding of a domain, being able to collectively manage them is a much-needed feature in the area of Semantic Web. This paper presents VocBench, a web platform for collaboratively managing Semantic Web datasets, and ShowVoc, an interface that facilitates the browsing and display of controlled vocabularies. The paper showcases how the European Science Vocabulary (EuroSciVoc), a taxonomy that represents fields of science, including life sciences and healthcare is managed in the tools in a collaborative way and exploiting Linked Open Data features.

2. The European Science Vocabulary (EuroSciVoc)

The Community Research and Development Information Service (CORDIS) is the European Commission's primary source of results from the projects funded by the EU's framework programmes for research and innovation, from FP1 to Horizon Europe. CORDIS has a rich and structured public repository with all project information held by the European Commission such as project factsheets, participants, reports, deliverables, and links to open-access publications.

To help users navigate and retrieve projects more easily, CORDIS has always needed to classify its projects based on their scientific field. This is why CORDIS developed a taxonomy called the European Science Vocabulary (EuroSciVoc) since taxonomies provide a standardized approach and improve data discoverability. The EuroSciVoc taxonomy represents all the main fields of science that were discovered from CORDIS content and organized through a semi-automatic process based on Natural Language Processing (NLP) algorithms. It contains more than 1000 categories in 6 languages (English, French, German, Italian, Polish and Spanish) and each category is enriched with relevant keywords extracted from the textual description of the projects. It is specifically developed as a reference vocabulary for the Open Science community and is aligned with Linked Open Data standards, as it is modelled in SKOS.

Reuse and interoperability are prominent features of EuroSciVoc. A notable example is its incorporation into the EU Knowledge Graph, particularly for classifying https://kohesio/ projects

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
*
aniko.gerencser@publications.europa.eu (A. Gerencsér); baya.remaoun@publications.europa.eu (B. Remaoun); enrico.bignotti@ext.publications.europa.eu (E. Bignotti)

– initiatives funded under the EU's Cohesion policy (2014-2020). EuroSciVoc was used to classify around 300 000 out of 2 million Kohesio projects (15%). These classifications will be added in the EU Knowledge Graph March 2025 release. In terms of interoperability, EuroSciVoc is aligned to other resources such as https://op.europa.eu/en/web/eu-vocabularies/concept-scheme/-/resource?uri=http://eurovoc.europa.eu/100141, https://agrovoc.fao.org/browse/agrovoc/en/ and to https://www.ncbi.nlm.nih.gov/mesh/ (through EuroVoc), therefore facilitating data integration across systems and domains and improved search capabilities.

Since CORDIS contains data about research projects, it must represent information about healthcare and life sciences projects. As of version 1.5, EuroSciVoc provides comprehensive coverage of such domains:

- 176 subcategories under "health and medicine"
- 222 categories in life sciences, encompassing "agricultural sciences", "biological sciences", and subcategories
- Classification of roughly 24,000 projects funded under H2020 and Horizon Europe:
  - 6,181 projects classified under "health and medicine" and its subcategories
  - 17,368 projects in life sciences categories

3. Managing and visualizing EuroSciVoc in VocBench and in ShowVoc

The Publications Office of the European Union provides corporate access to two complementary open-source platforms based on Semantic Web standards: https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/solution/vocbench and https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/solution/showvoc. These tools, supported by the Digital Europe Programme of the European Commission, create a versatile environment for RDF dataset management, with specialized features for ontologies, thesauri, taxonomies and lexicons.

VocBench serves as a collaborative web platform specifically designed for Semantic Web dataset development. The platform's strengths lie in its focus on collaboration, differentiated user roles, workflow management for content validation, and RESTful API integration through Semantic Turkey. The Publications Office of the EU offers https://op.europa.eu/en/web/eu-vocabularies/vocbench to VocBench for all EU institutions, bodies, and agencies, enabling semantic dataset managers to collaboratively edit their vocabularies. This centralized system offers two key advantages: collaborative editing with differentiated access levels and validation workflows, and enhanced Linked Data functionalities for project linking, alignment creation, and mapping validation.

ShowVoc complements VocBench by providing specialized visualization and dissemination features for RDF-based datasets. Based on the same semantic architecture as VocBench, ShowVoc provides an intuitive browsing interface and cross-dataset features such as global search and translation API. Additionally, the platform offers an enhanced visualization of alignments, offering both a per-dataset and a global view of display, including an expandable tree structure and a graph view. These tools enable users to analyze alignments across multiple datasets, providing an overview of their interconnections and relationships.

EuroSciVoc is one of over 400 vocabularies centrally managed within the corporate instance of VocBench of the Publications Office of the EU. VocBench and ShowVoc facilitate the collaborative editing and publication workflow of the taxonomy. Since EuroSciVoc undergoes continuous updates, including the addition of keywords to improve its classification features on the CORDIS website, VocBench enables the collaborative management of the categories of the taxonomy, seamless revision and validation workflow and the yearly release cycle. Once the taxonomy updates are finalized, EuroSciVoc is published in ShowVoc, which serves as a platform for both sharing data with users and enabling intuitive browsing of the taxonomy's categories and its alignments.

Declaration on Generative AI

The authors have not employed any Generative AI tools.