

# The EuroVoc Thesaurus: Management, Applications, and Future Directions

Lucy Walhain<sup>◇</sup>

Sébastien Albouze<sup>‡</sup>

Anikó Gerencsér<sup>◇</sup>

Mihai Paunescu<sup>†</sup>

Vassilis Tzouvaras<sup>†</sup>

Cosimo Palma<sup>\*</sup>

<sup>◇\*</sup> Publications Office of the European Union  
20 rue de Reims, L-2985 Luxembourg

<sup>†‡</sup> infeurope S.A.

113, rue Adolphe Fischer L-1521 Luxembourg

## Abstract

This paper provides a comprehensive overview of *EuroVoc*, the European Union’s multilingual thesaurus. The paper highlights *EuroVoc*’s significance in the legislative and publications domain, examining its applications in improving information retrieval systems and multi-label text classification methods. Various technological tools developed specifically for *EuroVoc* classification, including *JEX*, *PyEuroVoc*, and *KEVLAR*, are reviewed, demonstrating the evolution from basic classification systems to sophisticated neural architectures. Additionally, the paper addresses the management practices managing *EuroVoc*’s continuous updating and expansion through collaborative tools such as *VocBench*, emphasising the role of interinstitutional committees and specialised teams in maintaining the thesaurus’s accuracy and relevance. A substantial part of the paper is dedicated to *EuroVoc*’s alignment with other semantic resources like *Wikidata* and *UNESCO*, detailing the challenges and methodologies adopted to facilitate semantic interoperability across diverse information systems. Finally, the paper identifies future directions that include modular extensions of *EuroVoc*, federated models, linked data approaches, thematic hubs, selective integration, and collaborative governance frameworks.

## 1 Introduction

The European Union’s legislative framework encompasses a vast array of documents across multiple languages, necessitating robust systems for organisation and retrieval. *EuroVoc*<sup>1</sup> stands as a

cornerstone in this infrastructure as a comprehensive multilingual thesaurus specifically designed to systematise EU legislative documentation. Despite its instrumental role in numerous research experiments and practical applications, as evidenced by multiple published studies (see section 1.1), a comprehensive examination of *EuroVoc* as a foundational resource has remained notably absent from the literature. This paper addresses this significant gap by providing an in-depth analysis of *EuroVoc*’s structural features and functional capabilities. Furthermore, we examine the ongoing efforts within European Institutions to align *EuroVoc* with other semantic resources, enhancing its interoperability and utility.

### 1.1 A Review of Literature on EuroVoc

The literature concerning *EuroVoc* is extensive, reflecting both its historical and institutional significance. First published in 1984, *EuroVoc* was designed as a multilingual thesaurus to facilitate the indexing and retrieval of documents across the diverse linguistic landscape of European institutions. Since then, the Publications Office of the European Union has been responsible for updating and publishing *EuroVoc*. The thesaurus has evolved significantly over the past four decades.

This literature review examines the body of research surrounding *EuroVoc*, organised into two key dimensions: the technological tools and applications developed to leverage *EuroVoc*’s capabilities with a focus on the challenges and advancements in multi-label text classification using the *EuroVoc* framework; *EuroVoc*’s role as a linguistic and informational resource. For the sake of strictness, only contributions published after 2013 are considered here. This year marked a turning point, when the new linked data paradigm and the consequent interoperability standards led to changes in previous thesaurus modellings (Publications Office of the European Union, 2020).

<sup>◇</sup>name.surname@publications.europa.eu

<sup>†</sup>name.surname@ext.ec.europa.eu

<sup>‡</sup>name.surname@ext.publications.europa.eu

<sup>\*</sup>name.surname@ec.europa.eu

<sup>1</sup><https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc>

### 1.1.1 EuroVoc as a Resource for Legislation

In the legal domain, *EuroVoc* has proven particularly valuable for improving information retrieval systems beyond basic document classification. Cornoiu and Valean (Cornoiu and Valean, 2015) demonstrate *EuroVoc*'s effectiveness when integrated with Wikipedia knowledge bases and legal ontologies to create legal information retrieval mechanisms that bridge terminology gaps between legal professionals and laypeople.

Boella et al. (2013) established a foundation by developing one of the first comprehensive systems for multi-label classification of legislative texts into *EuroVoc* descriptors, based on the Support Vector Machine algorithm trained using the *JRC-Acquis corpus*<sup>2</sup>. Building upon this framework, Schmedding et al. (2018) expanded the application domain to European case law summarisation, demonstrating how *EuroVoc*-based classification could enhance accessibility and understanding of complex legal materials (see Figure 1 for an overview of annotation's comprehensiveness per year).

Addressing the multilingual challenges inherent in European legal systems, Gupta et al. (2012) pioneered cross-language similarity search techniques that leverage *EuroVoc* as a conceptual bridge across linguistic boundaries. The field advanced methodologically when Caled et al. (2022) introduced hierarchical label attention networks that exploit the intrinsic taxonomic structure of *EuroVoc* descriptors, substantially improving classification accuracy for legislative content.

Most recently, Bocchi and Palmero Aprosio (2024) challenged conventional approaches by examining the limitations of title-based classification for European laws, revealing that while document titles provide valuable signals, comprehensive content analysis remains essential for accurate *EuroVoc* multi-label classification.

Thanks to experiments conducted with *EuroVoc* and UNBIS<sup>3</sup> thesauri de Miranda Guedes and

<sup>2</sup>The JRC-Acquis corpus contains around 23,000 documents labeled with averagely six *EuroVoc* descriptors

<sup>3</sup>The UNBIS (United Nations Bibliographic Information System) Thesaurus is a multilingual controlled vocabulary created and maintained by the Dag Hammarskjöld Library of the United Nations Department of Public Information. It contains terminology used for subject analysis of documents and other materials relevant to United Nations programmes and activities, and is available in all six official UN languages. Source: <https://research.un.org/en/thesaurus>. UNBIS is considered *EuroVoc*'s closest conceptual counterpart in the international organization domain, and stands as the second most linguistically comprehensive thesaurus after *EuroVoc* in terms

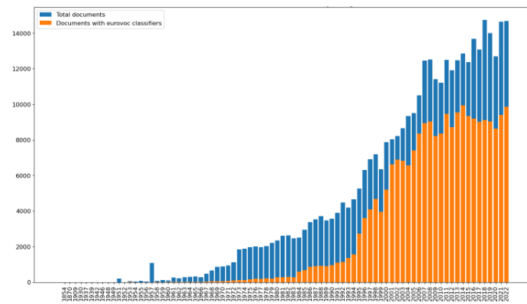


Figure 1: Number of documents per year (with percentage tagged with EuroVoc labels highlighted in orange) as in Bocchi et al. (2024).

Moura (2018) by examining in particular semantic warrant and cultural hospitality in multicultural contexts, it has been possible to understand how knowledge representation systems accommodate diverse cultural perspectives in legal and policy contexts.

### 1.1.2 Tools for EuroVoc-based Multi-label Text Classification

The proliferation of tools and resources specifically designed for *EuroVoc* classification represents a significant advancement in making multilingual legal document classification more accessible and efficient for researchers and practitioners. Steinberger et al. (2012) pioneered this movement with the *JRC EuroVoc Indexer* (JEX), a freely available multi-label categorisation tool that established essential benchmarks for automated *EuroVoc* classification and provided a foundation for subsequent developments. Building upon this foundation, Avram et al. (2021) introduced *PyEuroVoc*, a comprehensive Python-based toolkit that streamlined the implementation of multilingual legal document classification using *EuroVoc* descriptors, significantly lowering the technical barrier to entry for researchers working with diverse European languages.

Finally, Bocchi et al. (2024) unveiled *KEVLAR*, positioned as the complete resource for *EuroVoc* classification of legal documents, which consolidates previous advancements while introducing novel techniques and comprehensive datasets that address longstanding challenges in the field.

These tools collectively demonstrate the research community's commitment to developing accessible, efficient, and culturally nuanced approaches to *EuroVoc*-based classification, enabling broader

of language coverage, which further underscores *EuroVoc*'s preeminent status in the multilingual thesauri landscape.

adoption across various legal information systems while acknowledging the complex multilingual and multicultural dimensions of European legal and policy documentation.

## 1.2 Paper’s Contributions

Our systematic review of the literature surrounding *EuroVoc* reveals a significant gap in the existing research landscape.

The majority of published works have focused primarily on leveraging *EuroVoc* for Natural Language Processing applications, topic modelling methodologies, and information retrieval systems. These studies typically treat *EuroVoc* as a means to an end rather than as an object of study in its own right.

While these applications have undoubtedly advanced our understanding of how *EuroVoc* can enhance NLP tasks, they have not adequately addressed the fundamental semantic structure, ontological properties, and interoperability potential of the thesaurus itself. Despite the 2013 paradigm shift toward linked data principles in *EuroVoc*’s development, as noted in the *EuroVoc Handbook* (Publications Office of the European Union, 2020), relatively few studies have examined the implications of this transition for Semantic Web integration. The work of Paredes-Valverde et al. (Paredes et al., 2008) represents an early recognition of this potential, but comprehensive follow-up research exploring actual implementations of *EuroVoc* within the Semantic Web ecosystem remains sparse.

In particular, the research addressing ontological alignments between *EuroVoc* and other knowledge organisation systems, interoperability mechanisms across diverse EU information systems, and formal evaluations of *EuroVoc*’s compliance with contemporary linked data principles remain poor. Our paper addresses this research gap by providing a thorough examination of *EuroVoc* as a semantic resource, analysing its structural properties, ontological foundations and potential for alignment within the wider linked data ecosystem.

## 2 Management of EuroVoc

The effective management of *EuroVoc* is essential to maintaining its role as a comprehensive, multilingual thesaurus that supports the indexing and retrieval of EU-related documents. Originally created to process documentary information, *EuroVoc* has evolved to cover a wide range of domains. Its

management involves a structured approach to ensure that the thesaurus remains up-to-date, relevant, and accessible to users across the EU and beyond. *EuroVoc*’s structure and content supports precise classification and retrieval across thematic areas such as politics, law, and economics and its multilingual availability in all 24 official EU languages promotes cross-border information exchange.

The governance of the thesaurus involves multiple layers of collaboration and oversight, including an interinstitutional committee and a dedicated Reference Data Team. This team coordinates contributions, edits the thesaurus, and oversees its publication, ensuring adherence to international standards for terminology and thesaurus management. Furthermore, *EuroVoc*’s integration of Semantic Web technologies and alignment with the Simple Knowledge Organization System (SKOS) model underscores its adaptability and integration into modern digital information systems.

Table 1: *EuroVoc* Thesaurus in Numbers

Feature	Quantity
Hierarchical levels	8
Domains	21
Micro-thesauri	127
Preferred terms	7,000+
Languages	24
Total terms	678,000
Terms per language	24,000
Hierarchical rel.	10,000
Associative rel.	5,000
Non-preferred terms	12,000
Aligned knowledge bases	17
RDF triples	800,000+
Updates per year	3-4

*EuroVoc* publishes semi-annually across platforms like the *Cellar* semantic repository and the *EU Vocabularies* website, ensuring both machine- and human-readable access. By leveraging collaborative tools like *VocBench* (Stellato et al., 2015)<sup>4</sup> and engaging a diverse working group of professionals, *EuroVoc* remains a dynamic resource that continually adapts to new challenges and requirements. This section explores its management, content, structure, and publication processes, highlighting the collaborative and technological frameworks that support its ongoing development.

### 2.1 Content and Structure of EuroVoc

*EuroVoc* is a structured, multilingual thesaurus designed to support information retrieval, indexing, and semantic interoperability across EU institutions

<sup>4</sup><https://vocbench.uniroma2.it/doc/>

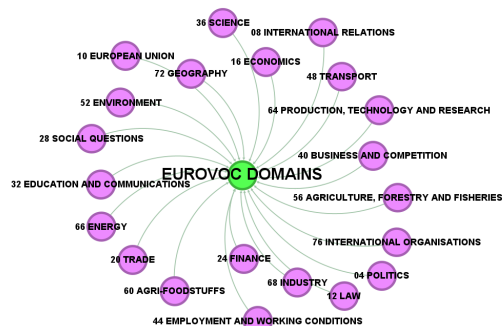


Figure 2: A glimpse of the 21 *EuroVoc* domains (all links are *skos:inScheme* relationships).

and external users. It is organised into 21 domains (see Figure 2), covering broad thematic areas such as politics, law, economics, science, and international relations, which align with EU policies and institutional activities. Each domain is further divided into 127 micro-thesauri (refer to Figure 3 for a better overview of domain’s sizes), which provide a finer level of classification by grouping related concepts or descriptors within specialised subcategories. For example, within the law domain, micro-thesauri cover areas like EU law, international law, and criminal law, while the economics domain includes micro-thesauri for financial markets, taxation, and economic policy. This hierarchical organisation ensures precise classification and retrieval of information across a wide range of EU-related topics (for an overview of these numbers, refer to Table 1).

*EuroVoc* is a multilingual thesaurus, available in 24 official EU languages, ensuring consistent terminology use across the European Union’s legislative, administrative, and research domains. In addition to the official EU languages, it also includes translations in languages of candidate countries and international partners, further expanding its reach and facilitating cross-border information exchange. Each concept maintains a unique identifier (URI), allowing for precise alignment of labels across different languages while preserving semantic integrity. This multilingual structure supports interoperability in a diverse linguistic environment, enabling efficient retrieval and classification of EU-related information for a global audience.

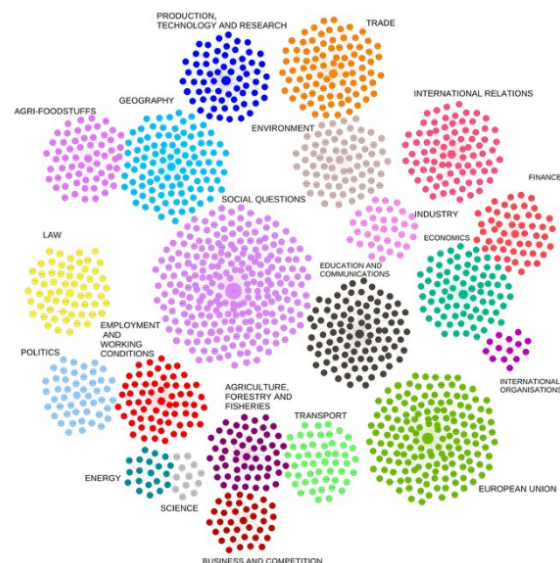


Figure 3: A representation of the *EuroVoc* Topic Clusters as in (Mahrouseh et al., 2022).

At the concept level, *EuroVoc* employs SKOS-based relationships to define structured connections between terms. It includes hierarchical relations, where a broader term (*skos:broader*) represents a more general concept and a narrower term (*skos:narrower*) denotes a more specific one. Additionally, associative relations (*skos:related*) link thematically connected concepts outside strict hierarchical structures. Each concept is also assigned a preferred term (*skos:prefLabel*), with alternative labels (*skos:altLabel*) capturing synonyms or variations to enhance searchability. These structured relationships ensure consistency in terminology across multilingual and multidisciplinary contexts.

To enhance interoperability, *EuroVoc* establishes SKOS-compliant mappings with external vocabularies and classification systems. These mappings facilitate semantic alignment and data exchange by linking *EuroVoc* concepts to equivalent or related terms in other knowledge organisation systems. Key SKOS mapping properties used in *EuroVoc* include *skos:exactMatch* for full equivalence and *skos:closeMatch* for near equivalence, while *skos:broadMatch* and *skos:narrowMatch* for hierarchical correspondences, and *skos:relatedMatch* for associative links might be considered in the future. *EuroVoc* is mapped to external resources such as UNBIS Thesaurus (United Nations), AGROVOC (Food and Agriculture Organization), GEMET (European Environment Agency), and national classification systems (see section 4 for further information on the topic).



## 2.2 Governance and Collection of Contributions

*EuroVoc* is governed by an inter-institutional committee which oversees its maintenance, update and biannual publication. The committee consists of members from various EU institutions, including the Council of the European Union, the European Parliament, and the Court of Justice, among others. The Reference Data Team, responsible for the maintenance and publication of various controlled vocabularies on *EU Vocabularies*, facilitates the work by analysing contributions, editing the thesaurus in *VocBench* and proceeding with its publication on *EU Vocabularies* and other platforms.

The possibility of contributing to *EuroVoc* is open to any user, either by completing the contribution form on *EU Vocabularies* or by directly contacting the Reference Data Team<sup>5</sup>. The contributions are then analysed, this process involves the compilation of potential new concepts (i.e. candidates), the identification of the corresponding domain and micro-thesaurus, and the addition of definitions that follow ISO standards on terminology work, and information and documentation. The list of candidates is then sent to a working group tasked with validating the thesaurus content. This working group comprises a diverse range of professionals, including terminologists, librarians, cataloguers, and knowledge managers.

Following the validation of the candidates' list by the working group, it is forwarded to the inter-institutional committee for final approval.

## 2.3 Collaborative Workflow in VocBench

The editorial work is performed in *VocBench*, where both the Reference Data Team and the working group work collaboratively. Candidates are added to the *EuroVoc* project<sup>6</sup> in *VocBench* in a candidate scheme (see Figure 4) accessible to all members.

Members are invited to add editorial notes (see Figure 5) to express their opinion on the proposed concept. These editorial notes remain internal and are never published in the thesaurus. They are used to facilitate collaboration and initiate discussion.

The Reference Data Team coordinates meet-

<sup>5</sup>[OPEUVOCABULARIES@OP-EU-VOCABULARIES@publications.europa.eu](mailto:OPEUVOCABULARIES@OP-EU-VOCABULARIES@publications.europa.eu)

<sup>6</sup>Each RDF-based dataset (ontology, thesaurus, taxonomy) in *VocBench* is called a "project". Links can be established between projects, enabling the creation of mappings between concepts stored in different projects.

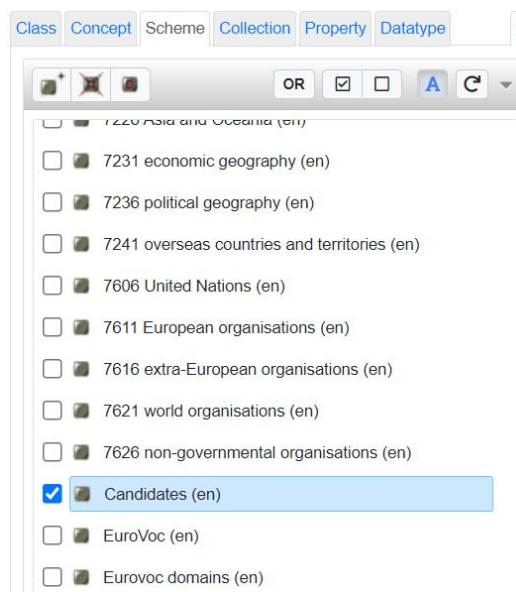


Figure 4: Candidates scheme in *VocBench*

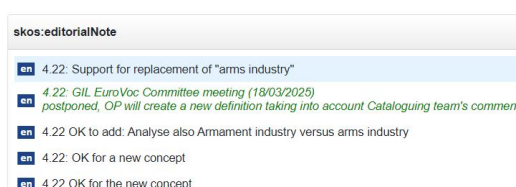


Figure 5: *EuroVoc* editorial notes in *VocBench*

ings with the working group to evaluate each candidate and their associated properties, such as domain, micro-thesaurus, definition, and related terms. When consensus is reached on a candidate, it is forwarded to the interinstitutional committee for final approval. If consensus is not achieved, the decision can be postponed for further clarification or the candidate may be rejected if deemed irrelevant to the thesaurus.

The collaborative approach to *EuroVoc*'s maintenance and update ensures its status as a robust and dynamic resource. By leveraging the expertise of diverse professionals and facilitating open contributions, the process enhances the thesaurus's accuracy and relevance. The structured framework for discussion and validation, supported by tools like *VocBench*, ensures thorough evaluation and efficient consensus-building. This method not only aligns *EuroVoc* with EU institutional needs but also ensures it remains adaptable to evolving terminological trends, maintaining its value and utility for a wide range of stakeholders.

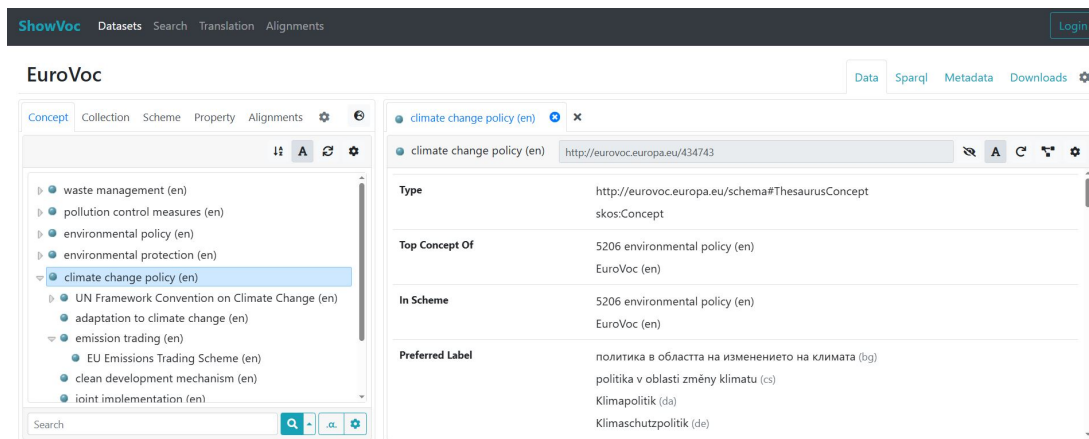


Figure 6: *EuroVoc* in ShowVoc

## 2.4 Translation and import to VocBench

The editorial process for *EuroVoc* begins with work conducted exclusively in English. Once the candidate terms and concepts have been thoroughly vetted and approved, they are forwarded to the European Commission Directorate-General for Translation. This team is tasked with translating all the elements of the candidates, including preferred labels, alternative labels, and definitions, into the 23 official languages of the European Union. Once the translation process is done, these multilingual components are imported back into *VocBench*, where they undergo a meticulous review by the Reference Data Team. They validate the translations to ensure they are both accurate and consistent across all languages, thus maintaining the high standards of quality and reliability that *EuroVoc* demands. This comprehensive translation and validation process is pivotal in ensuring that *EuroVoc* serves as an accessible and dependable resource for all EU Member States.

## 2.5 Publication of EuroVoc

*EuroVoc* is published on a semi-annual basis on multiple dissemination platforms, providing both machine- and human-readable access for its users:

- In the *Cellar* semantic repository<sup>7</sup> of the Publications Office, available for humans and machines via a SPARQL endpoint and allowing an API connection for systems;
- On the *EU Vocabularies* website<sup>8</sup> which offers a human-readable browsing experience in multiple views (tree view and alphabetical

list view) and download in various formats such as RDF, XML, MARC-XML, TBX and Excel;

- In *ShowVoc*<sup>9</sup>, a platform providing a user-friendly browsing interface for RDF-based controlled vocabularies. *ShowVoc* is based on the same semantic architecture as *VocBench* and offers an intuitive browsing interface and enhanced visualisation of alignments (see Figure 6). *ShowVoc* is also integrated to EU Vocabularies in the Advanced view of the dataset, expanding the browsing and visualisation experience on the website.
- Additionally, in multiple open data portals and reference data registries such as [data.europa.eu](https://data.europa.eu) or [bartoc.org](https://bartoc.org).

## 3 Use of EuroVoc

*EuroVoc* is used to categorise and index documents, from legislation to general publications and library resources, facilitating the organisation, the search and retrieval of information related to EU activities. It is used in various document management systems, databases and websites of EU institutions, in EU institutional and national government libraries as well as in academia and research institutes. In the following chapter we highlight a few use cases of *EuroVoc* in the Publications Office and in EU institutions.

### 3.1 EuroVoc in Eur-Lex

*EUR-Lex*<sup>10</sup> is an online portal that provides access to the European Union law and other docu-

<sup>7</sup><https://op.europa.eu/en/web/cellar>

<sup>8</sup><https://op.europa.eu/en/web/eu-vocabularies>

<sup>9</sup><https://showvoc.op.europa.eu/%23/home>

<sup>10</sup><https://eur-lex.europa.eu/homepage.html>



Figure 7: "Browse by subject" feature on the OP portal.

ments such as case-law and national law of Member States. *EuroVoc* is used to describe and index documents published in EUR-Lex. Each document is assigned with *EuroVoc* descriptors.

### 3.2 EuroVoc for Cataloguing

The Publications Office of the European Union is responsible for publishing and disseminating the publications of the EU institutions, agencies, and bodies. All published documents are accessible to the public.

Cataloguers assign *EuroVoc* descriptors to each document published, adding these to the subject metadata. This allows users to browse by subject using *EuroVoc* domains (see Figure 7), microthesauri, and concepts. *EuroVoc* organises its conceptual hierarchy across eight levels. Documents receive annotations with one or more concepts (descriptors), but typically exclude both ancestors and descendants of an assigned concept from the same document's annotation. Since *EuroVoc* is multilingual, each user can search in their language without affecting the search results.

Users can also search for documents using *Publio* (see Figure 8), an artificial intelligence tool that performs searches using keywords. *Publio* uses classifications and categories available on the Publications Office portal, such as *EuroVoc* descriptors.

### 3.3 Domain Classification in IATE: Integration with EuroVoc

The *InterActive Terminology for Europe* (IATE) is the European Union's multilingual terminology database, serving as the central repository for specialised terminology across all EU institutions. Launched in 2004, *IATE* facilitates translation consistency and linguistic precision in EU communications by providing standardised multilingual terminology across diverse subject domains.

**Domain Classification Framework** *IATE* implements the *EuroVoc* thesaurus as its principal domain classification system (*IATE Support Team*,

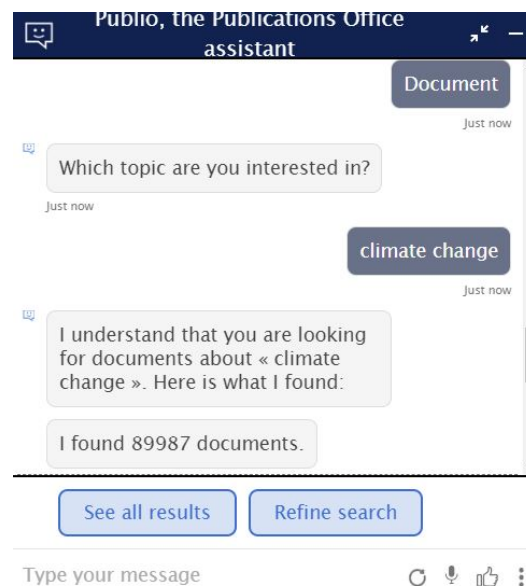


Figure 8: Publio, the Publications Office AI assistant

2023a). Each *IATE*<sup>11</sup> entry uses *EuroVoc* to indicate the domains to which the respective term belongs. Terminologists and translators have the flexibility to assign various *EuroVoc* descriptors to an entry, offering a comprehensive understanding of its domain associations. Currently, *IATE* utilises the complete *EuroVoc* thesaurus (version 4.6), marking a significant improvement over previous implementations that were restricted to only the first three hierarchical levels. The classification framework consists of:

- Primary domains (level one, identified by two-digit codes)
- Subdomains (level two, represented by four-digit codes)
- Descriptors (levels three through eight, without numerical identifiers)

**Specialised Legal Classification** To accommodate specialised legal terminology, *IATE* incorporates a secondary LAW branch (designated as '14 LAW') that integrates the classification system employed by the Court of Justice of the European Union (CJEU) (*IATE Support Team*, 2023a). Given that numerous subdomains within the CJEU LAW branch correspond to classifications already present in the *EuroVoc* thesaurus, the domain filtering functionality in both search interfaces and *IATE* data

<sup>11</sup>Record in the *IATE* database that typically contains terms, definitions, domains, etc.

exports automatically includes equivalent domains when available (IATE Support Team, 2023c).

**Domain Detection Functionality** The system offers automated domain detection capabilities, accessible through the full entry view interface (IATE Support Team, 2023b):

1. Users can access this feature by selecting the Domain label at the Language Independent Level (LIL) (IATE Support Team, 2023d)
2. The "Domain detection" option initiates a query to the Domain Classifier tool developed by the Joint Research Centre
3. The classifier generates *EuroVoc* classification recommendations based on entry content analysis
4. Users may select specific proposals or opt for higher-level domain categories (e.g., selecting the broader "Trade" category instead of the more specific "product quality" subdomain)

This integration of established taxonomies with specialised classification systems enables precise domain categorisation while maintaining terminological consistency across EU institutional communications.

## 4 Alignments between EuroVoc and other multilingual Vocabularies

Alignments establish correspondences between concepts, creating a comprehensive, interconnected knowledge ecosystem that improves information retrieval and multilingual access. For example, in Figure 11 a simple query over the Wikidata SPARQL-endpoint returns the alignments between *EuroVoc* and the UNESCO thesaurus. However, they imply significant challenges: they constantly evolve, requiring considerable maintenance, often lacking dedicated tools to assist in the process, as well as necessitating restarting alignment procedures from scratch when updates occur. In the following paragraphs we show how these challenges are addressed at the Publication Office of the European Institutions.

### 4.1 Aligning EuroVoc with Wikidata: Challenges and Approaches

The alignment of *EuroVoc* with Wikidata<sup>12</sup> has historically been a complex endeavour due to the

<sup>12</sup>[https://www.wikidata.org/wiki/Wikidata:Main\\_Page](https://www.wikidata.org/wiki/Wikidata:Main_Page)

```
SELECT
  ?item ?itemLabel ?eurovocid ?eurovocuri
WHERE
{
  ?item wdt:P5437 ?eurovocid
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
  BIND(uri(CONCAT('http://eurovoc.europa.eu/', ?eurovocid)) AS ?eurovocuri)
}
```

Figure 9: Wikidata query for EuroVoc-ID property statement.

intrinsic characteristics of both datasets. Their large size, multilingual nature, and extensive sets of alternative labels rendered a brute-force approach impractical. Additionally, the continuous evolution of both resources—through the addition, modification, and deprecation of concepts and labels—further complicated the process. Several automated methodologies were tested, including a scripting-based approach (*Python*<sup>13</sup>) and an ETL-based<sup>14</sup> solution (*LinkedPipes*<sup>15</sup>).

However, both approaches proved unsatisfactory due to technical limitations, particularly inconsistencies in responses from the Wikidata- SPARQL endpoint when handling large query volumes (see Figure 9). Moreover, the resulting mappings required post-processing validation to ensure quality. In practice, a hybrid approach leveraging *OpenRefine*<sup>16</sup> was adopted (see Figure 10), enabling editors to interact directly with the alignment process and integrate validation without additional tools. Nevertheless, incorporating *OpenRefine* into a sustainable workflow for both initial alignment and ongoing maintenance was considered impractical: its use was put on hold while alternative solutions continue to be explored.

### 4.2 Integrating EuroVoc Alignments into the Editorial Workflow: current work-in-progress

**Historical Approaches and Legacy Tools** Historically, due to its extensive size and broad thematic coverage, the alignment of *EuroVoc* with other vocabularies has been conducted outside the

<sup>13</sup><https://www.python.org/>

<sup>14</sup>Extract, transform, load (ETL) is a three-phase computing process where data is extracted from an input source, transformed (including cleaning), and loaded into an output data container. The data can be collected from one or more sources and it can also be output to one or more destinations (Source: Wikipedia).

<sup>15</sup><https://etl.linkedpipes.com/>

<sup>16</sup>OpenRefine is an open-source desktop application for data cleaning, transformation, and enrichment. It provides tools to explore, clean, reconcile, and enhance data without programming knowledge. Further information can be retrieved at <https://openrefine.org/>



```

<rdf:description rdf:about="http://eurovoc.europa.eu/1048">
  <rdftype rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#about"/>
  <closeMatch xmlns="http://www.w3.org/2004/02/skos/core#" rdf:resource="http://www.wikidata.org/entity/Q11771944"/>
</rdf:description>

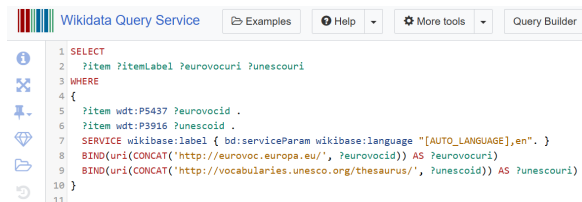
<rdf:description rdf:about="http://eurovoc.europa.eu/1047">
  <rdftype rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#about"/>
  <closeMatch xmlns="http://www.w3.org/2004/02/skos/core#" rdf:resource="http://www.wikidata.org/entity/Q1628947"/>
  <closeMatch xmlns="http://www.w3.org/2004/02/skos/core#" rdf:resource="http://www.wikidata.org/entity/Q122238"/>
</rdf:description>

<rdf:description rdf:about="http://eurovoc.europa.eu/105">
  <rdftype rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#about"/>
  <closeMatch xmlns="http://www.w3.org/2004/02/skos/core#" rdf:resource="http://www.wikidata.org/entity/Q10903133"/>
  <closeMatch xmlns="http://www.w3.org/2004/02/skos/core#" rdf:resource="http://www.wikidata.org/entity/Q38566"/>
</rdf:description>

<rdf:description rdf:about="http://eurovoc.europa.eu/1051">
  <rdftype rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#about"/>
  <closeMatch xmlns="http://www.w3.org/2004/02/skos/core#" rdf:resource="http://www.wikidata.org/entity/Q157031"/>
</rdf:description>

```

Figure 10: A screenshot of the RDF file resulting from the *OpenRefine* workflow.



```

1 SELECT
2   ?item ?itemLabel ?eurovocuri ?unescuri
3 WHERE
4 {
5   ?item wdt:P5437 ?eurovocid .
6   ?item wdt:P3916 ?unescoid .
7   SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
8   BIND(uri(CONCAT("http://eurovoc.europa.eu/", ?eurovocid)) AS ?eurovocuri)
9   BIND(uri(CONCAT("http://vocabularies.unesco.org/thesaurus/", ?unescoid)) AS ?unescuri)
10 }
11

```

Figure 11: Query on Wikidata SPARQL-endpoint showing alignment between Unesco and *EuroVoc* ontologies.

standard editorial workflow. These alignments were typically managed as separate tasks, independent of the routine maintenance and development of the thesaurus. However, there is a growing trend toward integrating alignment activities more closely into core operations by embedding them within a structured workflow based on commonly used tools. Initially, *EuroVoc* alignments were performed using tools such as *Silk Workbench* and *OpenRefine*, which provided manual and semi-automated methods for linking concepts across vocabularies. While effective in certain contexts, these tools required significant human intervention, lacked seamless integration with *EuroVoc*'s editorial environment, and were not fully optimised for ongoing maintenance and revision.

### Integrated VocBench-GENOMA Framework

A significant shift in the alignment process is now underway, transitioning toward the adoption of *GENOMA* (GENeric Ontology Matching Architecture)(Enea et al., 2015), which is integrated within *VocBench*. This transition marks a pivotal advancement in the way *EuroVoc* alignments are conducted. The *VocBench*-based workflow introduces several advantages, most notably the direct integration with the semantic repository where *EuroVoc* is maintained. This integration ensures that alignments are not only more systematic but also remain dynamically linked to updates within the thesaurus. Another major benefit of this workflow is the increased automation of the matching process, significantly reducing manual effort. Within this framework, *EuroVoc*

and the target vocabulary are incorporated into a dedicated alignment project, where the user is presented with structured lists of concepts ready for alignment. The user can either perform manual searches to establish semantic relations between corresponding concepts or initiate automated alignment tasks for specific lexicalisations. These tasks are executed under the Alignment Validation function of *VocBench*, which subsequently presents the user with the detected matches, allowing them to review and validate results individually or based on predefined criteria.

**Challenges in Multilingual Alignment** While automation enhances efficiency, it also introduces challenges, particularly due to the multilingual nature of *EuroVoc*. Variations in word meanings and structural differences between languages can lead to discrepancies in alignment results. The interpretation of terms across different vocabularies, especially in language pairs with significant semantic divergence, presents a potential risk of misalignment. Despite these challenges, the native multilingual capabilities of *VocBench*, coupled with its structured validation mechanism, provide a robust framework for managing these complexities.

## 5 Conclusions and Future Directions

This paper has provided an extensive review and analysis of *EuroVoc*, the multilingual thesaurus employed by the European Union for annotating EU documents. It has detailed *EuroVoc*'s structural characteristics, management practices, and its importance in enhancing legislative document retrieval and multi-label text classification. The paper examined various technological tools such as *JEX*, *PyEuroVoc*, and *KEVLAR* that demonstrate *EuroVoc*'s evolution and increasing sophistication. Moreover, it addressed significant challenges and strategies involved in maintaining semantic interoperability with other resources like Wikidata and UNESCO. Through this comprehensive analysis, the paper underscored *EuroVoc*'s role as a critical semantic resource within EU institutions.

**Future Work** Future directions should explore several innovative and practical approaches to enhance *EuroVoc*'s effectiveness, adaptability, and sustainability. Drawing from scenarios analysed in relation to EU agencies' specific needs, modular extensions and federated models could offer valuable frameworks that allow for flexibility, scalability,

and semantic consistency.

Linked data strategies and thematic hubs could further optimise semantic interoperability without overwhelming the core thesaurus structure. Selective integration methods would help maintain targeted growth, ensuring *EuroVoc* remains concise and relevant. To advance these strategic directions, collaborative governance frameworks should be considered to distribute maintenance responsibilities effectively among EU agencies.

Dynamic concept expansion, leveraging advanced AI-driven semantic elicitation tools, could automate and refine the identification and integration of relevant emerging concepts. Controlled vocabulary sets provided by agencies might also offer a structured yet flexible means of expanding *EuroVoc*'s coverage without compromising manageability. Additionally, further integrating advanced artificial intelligence methods could significantly boost *EuroVoc*'s utility and operational efficiency. Employing natural language processing and large language models could greatly enhance multilingual semantic tagging, classification accuracy, and automated ontology alignment. AI-driven analytics could proactively identify emerging concepts and semantic shifts, ensuring *EuroVoc* remains current and responsive to evolving legislative language and domains. Moreover, using AI-powered recommender systems could personalise user interactions, streamline content discovery, and improve overall user satisfaction.

Beyond these documented scenarios, additional recommendations include adopting advanced machine learning techniques for automated multilingual translations, quality control, and conflict resolution, thus addressing the semantic warrant challenges identified by [de Miranda Guedes and Moura \(2018\)](#). Enhancing user interfaces with intuitive search functionalities and adaptive visualisations would improve end-user experiences, facilitating easier navigation of an expanded thesaurus. Implementing robust version control and dependency tracking within VocBench would enhance management capabilities.

Finally, regular stakeholder training programs and feedback mechanisms could ensure that *EuroVoc* continues to evolve in alignment with the evolving informational landscape and user requirements across the EU institutions, extending the foundation established by [Bocchi et al. \(2024\)](#).

## 6 Acknowledgments

The authors express their sincere gratitude to the GIL EuroVoc Committee members who contribute to the maintenance and publication of *EuroVoc*, and the IATE Support and Development team for providing substantial information about the integration of *EuroVoc* in IATE.

We extend our thanks to Prof. Armando Stellato for his suggestions in shaping the final outline of the paper.

Finally, the authors thankfully acknowledge Denis Dechandon for his thorough review of the manuscript and for sharing his extensive knowledge of *EuroVoc*'s development history and operational aspects. His suggestions significantly improved the quality and accuracy of this work.

## References

- Andrei-Marius Avram, Vasile Pais, and Dan Ioan Tufis. 2021. [PyEuroVoc: A tool for multilingual legal document classification with EuroVoc descriptors](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 92–101, Held Online. INCOMA Ltd.
- Lorenzo Bocchi, Camilla Casula, and Alessio Palmero Aprosio. 2024. KEVLAR: The complete resource for EuroVoc classification of legal documents. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy.
- Lorenzo Bocchi and Alessio Palmero Aprosio. 2024. Title is (Not) all you need for EuroVoc multi-label classification of European laws. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy.
- Guido Boella, Luigi Di Caro, Daniele Rispoli, and Livio Robaldo. 2013. [A system for classifying multi-label text into eurovoc](#). In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, ICAIL '13*, page 239–240, New York, NY, USA. Association for Computing Machinery.
- Danielle Caled, Mário J Silva, Bruno Martins, and Miguel Won. 2022. Multi-label classification of legislative contents with hierarchical label attention networks. *International Journal on Digital Libraries*, pages 1–14.
- Sorina Cornoio and Honoriu Valean. 2015. Improving legal information retrieval using the wikipedia knowledge base, legal ontology and the eurovoc thesaurus. In *2015 19th International Conference on System Theory, Control and Computing (ICSTCC)*, pages 111–116. IEEE.
- Roger de Miranda Guedes and Maria Aparecida Moura. 2018. Semantic warrant, cultural hospitality and knowledge representation in multicultural contexts: experiments with the use of the eurovoc and unbis thesauri. *Advances in Knowledge Organization*, 16:442–449.
- Roberto Enea, Maria Teresa Pazienza, and Andrea Turbati. 2015. Genoma: Generic ontology matching architecture. In *AI\*IA 2015 Advances in Artificial Intelligence*, pages 303–315, Cham. Springer International Publishing.
- Parth Gupta, Alberto Barrón-Cedeno, and Paolo Rosso. 2012. Cross-language high similarity search using a conceptual thesaurus. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics: Third International Conference of the CLEF Initiative, CLEF 2012, Rome, Italy, September 17-20, 2012. Proceedings 3*, pages 67–75. Springer.
- IATE Support Team. 2023a. [Domains – iate online help](#).
- IATE Support Team. 2023b. [Experimental features \(\\*\) – iate online help](#).
- IATE Support Team. 2023c. [Filters – iate online help](#).
- IATE Support Team. 2023d. [Language-independent level – iate online help](#).
- Nour Mahrouseh, Szabolcs Lovas, Diana Njuguna, Noel Nellamkuzhi, Carlos Alexandre Soares Andrade, Wilhelmina Sackey, Anggi Irawan, and Orsolya Varga. 2022. [How the european union legislations are tackling the burden of diabetes mellitus: A legal surveillance study](#). *Frontiers in Public Health*, 10.
- Luis Polo Paredes, JM Álvarez Rodríguez, and Emilio Rubiera Azcona. 2008. Promoting government controlled vocabularies for the semantic web: the eurovoc thesaurus and the cpv product classification system. In *Proceedings of the 1st International Workshop on Semantic Interoperability in the European Digital Library (SIEDL 2008)*, pages 111–122.
- Publications Office of the European Union. 2020. [Eurovoc maintenance, publication and development handbook](#). Technical report, Publications Office of the European Union, Luxembourg.
- Florian Schmedding, Peter Klügl, David Baehrens, Christian Simon, Kai Simon, and Katrin Tomanek. 2018. Eurovoc-based summarization of european case law. In *AI Approaches to the Complexity of Legal Systems: AICOL International Workshops 2015-2017: AICOL-VI@ JURIX 2015, AICOL-VII@ EKAW 2016, AICOL-VIII@ JURIX 2016, AICOL-IX@ ICAIL 2017, and AICOL-X@ JURIX 2017, Revised Selected Papers 6*, pages 205–219. Springer.
- Ralf Steinberger, Mohamed Ebrahim, and Marco Turchi. 2012. [JRC eurovoc indexer JEX - a freely available multi-label categorisation tool](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 798–805, Istanbul, Turkey. European Language Resources Association (ELRA).
- Armando Stellato, Sachit Rajbhandari, Andrea Turbati, Manuel Fiorelli, Caterina Caracciolo, Tiziano Lorenzetti, Johannes Keizer, and Maria Teresa Pazienza. 2015. Vocbench: a web application for collaborative development of multilingual thesauri. In *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31–June 4, 2015. Proceedings 12*, pages 38–53. Springer.