Terminology: Towards a Systematic Integration of Semantics and Metadata

Denis Dechandon
Anikó Gerencsér
Publications Office of the European Union
denis.dechandon@publications.europa.eu
aniko.gerencser@publications.europa.eu

Maria Recort Ruiz
International Labour Office
recortruiz@ilo.org

Abstract

Knowledge sharing is one of the greatest challenges in businesses, public administrations and international organisations. Sharing large quantities of data to be easily used by IT applications and systems is essential to improve working methods, increase productivity, craft policy and implement decisions at institutional level.

This paper presents the first stage of a collaborative project undertaken between the Publications Office of the European Union (OP) and the International Labour Office (ILO) to align and map some of their terminology assets and controlled vocabularies (EuroVoc and the ILO thesaurus, taxonomy and 'gig' economy glossary). The project's objectives are to improve the quality of selected vocabularies, improve data accessibility, implement linked data and mappings and enrich terminology assets in order to move to a systematic use of semantics all through the authoring-translating-publishing chain and to support the building of knowledge graphs and the use of artificial intelligence at a later stage.

1. Introduction

Activities of businesses and public administrations increasingly depend upon internet data reused by IT applications and systems able to read and regurgitate semantic content, the ultimate purpose being to maximise the reuse of data for better decision- and policymaking.

While the semantic web, or web 3.0, allows anyone to say something about any topic, this "something" being combined and combinable with data from other sources, the main purpose of semantics[1] is to lead to knowledge extraction from enormous sets of raw data in various formats. This facilitates faster and more cost-effective access to meaningful and accurate data, to analyse it and convert it into knowledge. In this framework and further to the enabling of the communication between computers and the implementation of process automations, controlled vocabularies find a new rationale: supporting the disambiguation of content and improving and easing the larger implementation of technologies, such as automated tagging or machine translation.

In this landscape, new technologies and standards (e.g. RDF, RDFa, RDFS, OWL, SKOS, SPARQL) help *inter alia* infer, relate, interpret, and classify the implicit meanings of digital contents, and ensure the information's availability that citizens, decision-makers, linguists, researchers, etc. need. Furthermore, it relies on a shared understanding, that this information should be discovered and retrieved, even split in pieces between computers and servers. Thus,

[1] applied languages expressing interrelations of data in machine-readable format.

the objective is to get the right data to the expected place and allow intelligent applications to read and reuse the unstructured content created in the web 1.0, 2.0 and even 3.0, because of the standards and formats used, and missing or "wild" metadata.

Improving computer work, developing systems supporting trusted interactions over the web of data, enhancing and implementing "controlled vocabularies" created by documentalists, taxonomists, information scientists, etc. and the links between them and terminology assets developed by another professional community are some of the numerous topics we propose to address through a concrete collaborative project around the EuroVoc thesaurus, maintained by OP, and the ILO thesaurus, taxonomy and "gig" glossary. Hence, we will focus on:

- Already existing interweaving of terminology assets and controlled vocabularies,
- Various results and advantages of implementing technologies and standards mentioned above, associated with assets used by linguists and knowledge and information management professionals,
- Current achievements,
- Further steps to fit into the paradigm of the web 3.0,

to create assets maintained and used by linguists benefiting from an increased return on investment and to come to systematic use of semantics leading notably to redefine linguists' tools.

2. Controlled vocabularies

A. "Controlled" vocabularies?

International organisations, EU institutions and bodies, their respective Member States and their stakeholders, create, maintain, manage and implement various kinds of controlled vocabularies, like glossaries, classifications, taxonomies and thesauri.

Mostly used by different professional communities:

- Glossaries, dictionaries, traditional language thesauri and terminology databases are among the tools used when translating, interpreting or getting the basic meaning of a word, a term or a concept,
- Authority lists, classifications, ontologies, taxonomies and thesauri containing standardised linguistic contents (labels associated to identifiers) and keywords used by information specialists, knowledge managers, data scientists, etc. for:
    - Indexing /structuring contents,
    - Enhancing information discoverability on websites,
    - Exploiting data thanks to various technologies,
- Sets of technical metadata used by IT and semantic technology specialists in IT applications and for machine-to-machine communication.

Considering vocabularies and their respective purposes, we have the following taxonomy of knowledge organisation systems (KOS)[2]:

[2] Source: Zeng, Marcia. (2008). Knowledge Organization Systems (KOS). Knowledge Organization. 35. 160-182. 10.5771/0943-7444-2008-2-3-160

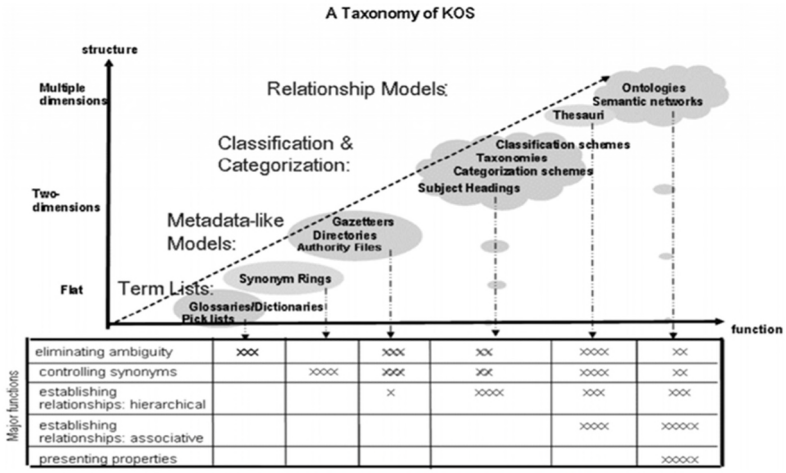


Figure 1. *An overview of the structures and functions of KOS*

Some might be considered as terminology assets used by linguists but all are KOS including "all types mentioned above, plus categorization schemes, classification schemes, dictionaries, gazetteers, glossaries, ontologies, semantic networks, subject heading schemes, and terminologies"[3].

Recently, sophisticated KOS appeared, **knowledge graphs**:

- "Knowledge graphs can be considered ontologies and more. [...] "A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge.""[4].

All KOS contribute to semantic, technical and terminological interoperability. Additionally, ontologies and data modelling, both required by the semantic web, strongly relate to future applications of knowledge management or artificial intelligence.

B. Making controlled vocabularies multilingual

The more accessing information tends to be global, the more KOS are made multilingual and support the provision of an extended access to contents available on the web (access to information is provided, whatever the language of each user, if the considered multilingual KOS covers it).

3. Tools and standards to create and maintain vocabularies

A. Terminology and controlled vocabularies

Terminology or controlled vocabulary management are two interrelated fields with different backgrounds and objectives complementing each other as regards knowledge sharing.

There is no single definition for terminology management or controlled vocabulary management. Terminology can be defined as "science studying the structure, formation, development, usage and management of terminologies in various subject fields" (ISO 1087-1) and its application as a "set of designations belonging to one special language" (ISO 1087-1). Terminology management involves planning and defining terminology processes, implementing computer aided terminology tools and regular updating of terminology assets. It deals with the capturing, processing, updating and preparation of terminological data.[5]

[3] https://www.hedden-information.com/taxonomies-as-knowledge-organization-systems/
[4] https://www.hedden-information.com/knowledge-graphs-and-ontologies/
[5] Hendrik J. Kockaert, Frieda Steurs. Handbook of Terminology. *Terminology tools.* Frieda Steurs, Ken de Wachter, Evy de Malsche. John Benjamins Publishing Company, Amsterdam, 2015.

While terminologies are more oriented for translation purposes (they include e.g. nouns and phrases including verbs or adjectives, part of speech, use, definitions), taxonomies and thesauri are mostly used for indexing, tagging or classifying purposes and constructed on a hierarchical basis establishing relationships between terms and concepts (they interrelate terms and information retrieval).

Data, metadata, technologies and standards will remain. KOS implemented in this new landscape will evolve. Improving, interlinking and using them will further give structure and meaning to the vast amount of unstructured data available in the cloud.

B. Standards

1. Terminology ISO standards

ISO standards provide "rules, guidelines or characteristics for activities or for their results, aimed at achieving the optimum degree of order in a given context"[6]. The European Parliament's Terminology Coordination Unit of the Directorate-General for Translation collected numerous relevant standards for this field[7].

2. Standards for controlled vocabularies8

The ISO 25964 standard, "Information and documentation - Thesauri and interoperability with other vocabularies" comprises:

- Part 1 (ISO 25964-1:2011): *Thesauri for information retrieval*, which "covers the development and maintenance of thesauri, both monolingual and multilingual, including formats and protocols for data exchange",
- Part 2 (ISO 25964-2:2013): *Interoperability with other vocabularies*, which "covers interoperability between different thesauri and with other types of structured vocabulary [...] not previously covered in any International Standard".

Additionally, standards defined by the W3C[9] are of utmost importance.

C. Data modelling

Most EU institutions and bodies dedicate efforts to creating ontologies or knowledge graphs and implementing data modelling as required for the semantic web, considering knowledge management and artificial intelligence.

Data models may have various meanings, depending on the chosen approach, on professionals working on them or on their objectives and purposes (models for business information, solution data, and physical data)[10].

4. Building bridges between controlled vocabularies

Interoperability is enabling at least two "systems or components to exchange information and to use the information that has been exchanged"[11].

[6] https://www.iso.org/deliverables-all.html
[7] https://www.iso.org/obp/ui/#iso:std:iso:704:ed-3:v1:en
[8] https://www.niso.org/standards-committees/iso-25964, https://www.iso.org/obp/ui/#iso:std:iso:25964:-1:ed-1:v1:en, https://www.iso.org/obp/ui/#iso:std:iso:25964:-2:ed-1:v1:en
[9] https://www.w3.org/standards/semanticweb/
[10] https://www.dataversity.net/state-art-data-modeling/#
[11] ISO 25964-1:2011 standard

A. Aligning controlled vocabularies and mapping concepts

Mappings aim to show relationships between concepts in different vocabularies. Including mappings to other vocabularies into controlled vocabularies supports interoperability[12]. This is valid for all vocabulary types, regardless of how concepts are represented:

Type of vocabulary	Concepts representation
Classification	Notations
Name authority list	Names
Ontology	Labels
Taxonomy	Category labels or notations
Terminology	Terms (or else, according to standards)
Thesaurus	Preferred terms

After mapping, improved KOS enhance the retrieval of content from the web.

As the semantic web is rapidly expanding, more and more KOS are made available and aligned to each other worldwide. This brings an expanding number of mappings between concept representations, as described in the ISO 29654-2:2013 standard, and a growing volume of linked (open) data.

However, mapping concepts implies dealing with some challenges:

- Identifying KOS worth being aligned,
- Having all KOS in processable formats by IT tools or human beings,
- Identifying technologies to use,
- Defining the mapping direction,
- Identifying elements to be mapped, depending on knowledge management systems,
- Setting mapping types to be defined, introduced and implemented,
- Making validated all suggested mappings by both taxonomists/information and knowledge management specialists in charge of selected KOS,
- Defining which KOS language versions should be aligned to each other,
- Ensuring mapping precision, as retrieving pre-indexed content is at stake,
- Having multilingual mappings checked by linguists and by taxonomists/information and knowledge management specialists,
- Defining and applying a governance (content and technical evolutions, ownership, responsibilities, roles, concept representations (lexicalisations for human beings and URIs for computers) among others).

Quite often, only a concepts subset included in a KOS are mapped with another one's concepts. A further step would envisage merging these KOS and/or considering one as an extension of the other.

B. Merging KOS

Different KOS compatible in structure, purpose and scope are easier to merge. However, to ensure integrating them coherently, the merging shall be handled with discipline, caring of eliminating duplicates, replacing terms, refining concepts, definitions or scope notes. *In fine*, the merged KOS will be enriched and consolidated, enhancing interoperability. This also applies when incorporating valuable discontinued KOS into another actively maintained KOS.

[12] ISO 25964-2:2013 standard

If mapping activity requires precision, merging does not, since a new KOS will be produced.

Before merging KOS, some questions should be answered:

- Are both KOS on the same subject?
- Can the target dataset integrate the whole source one? If not, which part can be incorporated? Would the left-out part become an extension of the target dataset?
- For merged but not top concept, which relationships are needed between it and the existing concepts in the enriched KOS?

After mapping concepts and merging KOS, remaining concepts can be considered as extensions of the improved KOS.

C. Defining extensions

Aligned with various other KOS, the EuroVoc thesaurus contains mappings defined as exact (skos:exactMatch) or inexact (skos:closeMatch).

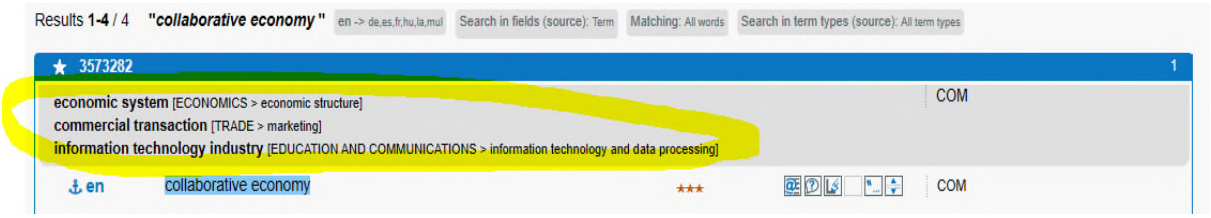
Further matches can be considered, like broad (skos:broadMatch) and narrow (skos:narrowMatch) matches.

In a given domain, matches can be defined between concepts in KOS of different granularities.

D. Pushing over the edge

1. EuroVoc and IATE (InterActive Terminology for Europe), the EU's terminological database

As IATE uses EUROVOC as its domain classification system, there is neither mapping nor merging. It rather supports the disambiguation of concepts.



A controlled use of EuroVoc descriptors in IATE would turn the database into a wonderful source of concepts and lexicalisations.

2. EuroVoc and LYNX (Legal Knowledge Graph for Multilingual Compliance Services)

The LYNX project aims to provide more effective ways of accessing huge amounts of digital regulatory compliance documents in the legal domain. After extracting the terms from the legal corpora and creating language resources from the results, the extracted data will be transformed into machine-readable formats and linked to additional legal resources such as EuroVoc, BabelNet or DBPedia[13]. The project foresees to create an ecosystem of smart cloud services based on a legal knowledge graph integrating and linking heterogeneous compliance data sources.

[13] Patricia Martín-Chozas, Víctor Rodríguez-Doncel: *Towards a Linked Open Data Cloud of Language Resources in the Legal Domain.* Law via the Internet Conference, Florence, 11-12 October 2018.

5. Four practical examples (EU-EuroVoc vs ILO-Thesaurus, Taxonomy, Glossary)

The European Commission, as all EU institutions and bodies creates, maintains, manages and implements various kinds of controlled vocabularies, like classifications, taxonomies and thesauri, which are mostly:

- Standardised labels and keywords used for content indexing or enhancing information discoverability on websites,
- Technical metadata sets used in IT applications.

International organisations within the United Nations system have different terminology and controlled vocabularies and taxonomy databases. The ILO manages three main databases: ILOTERM, the ILO Thesaurus and the ILO Taxonomy.

A. Working with some different assets created for various purposes

1. The Publications Office of the European Union

OP mainly publishes European Union institutions' publications (Decision 2009/496/EC, Euratom). Its core activities include producing and disseminating legal and general publications in paper or electronic formats like the daily Official Journal of the European Union in 23 (24 with Irish) official EU languages, ensuring long-term preservation of content produced by EU institutions and bodies and managing several websites providing digital access to information and data from the EU or offering specific services, tools and manuals.

a) EuroVoc, a multilingual, multidisciplinary thesaurus covering the activities of the EU

Among those tools, EuroVoc, a multidisciplinary thesaurus originally aggregated for processing EU institutions' documentary information, covers fields sufficiently wide-ranging for both Community and national points of view, emphasising parliamentary activities. EuroVoc is used outside the EU institutions, particularly by national and regional parliaments.

The thesaurus aims to provide a coherent indexing tool to manage efficiently documentary resources and execute documentary searches using KOS.

Containing terms in 24 EU languages, plus 3 EU accession candidate countries' languages (Macedonian, Albanese and Serbian), it currently counts 21 domains and 127 microthesauri.

OP maintenance team continuously adapts it considering developments in EU institutions' fields of activity, changes in its language arrangements and collecting and examining the proposals submitted by users through a form available on the EU Vocabularies website. A governance and a maintenance system were established to deal with those matters.

b) Implementation of the Simple Knowledge Organisation System (SKOS)

The EuroVoc maintenance was migrated from Excel to **VocBench2** with its 4.4 release, enabling the use of semantic web technologies to reflect W3C recommendations and latest trends in thesaurus standards.

The latest **VocBench3** offers a powerful editing environment with facilities for OWL ontologies, SKOS/SKOS-XL thesauri, OntoLex lexicons and any sort of RDF dataset management. Aiming to set new standards for flexibility, openness and expressive power, this RDF modelling platform is free and open source. OP manages its further development.

OP uses VocBench3 to maintain numerous controlled vocabularies and shares it increasingly with other institutions and organisations.

EuroVoc resources are modelled as direct extensions of the SKOS and SKOS-XL (SKOS eXtension for Labels (appendix B)) classes and properties. Some Dublin Core properties were reused.

2. The International Labour Organization

Founded in 1919, the Organization became the first specialized agency of the United Nations (UN) in 1946. It has a unique tripartite structure within the United Nations system bringing together governments, employers and workers to set labour standards, develop policies and devise programmes promoting decent work for all women and men.

Based in Geneva (Switzerland), ILO is the permanent secretariat of the International Labour Organization.

a) A terminology asset: The "gig" economy glossary

ILOTERM, public multilingual terminology database of ILO, contains more than 24,000 entries and 100,000 terms in seven official working languages of the Office. Most are related to labour issues, human rights and other UN related fields. Managed and regularly updated in close collaboration with technical experts by the Official Meetings, Documentation and Relations Department, ILOTERM is also used by linguists.

The "gig" economy glossary was prepared in the framework of research about non-standard forms of employment. This relatively new field required important work amounts from ILO's language specialists and technical experts.

The final glossary is available in Excel for internal use, while some of terms were included in ILOTERM. Often updated as this field constantly evolves, it is used by editors, translators, revisers, formatting operators and experts, and external users through ILOTERM.

b) ILO Taxonomy and Thesaurus

The ILO Taxonomy and Thesaurus provide (online) a list for web content subject tagging within the ILO.

Reflecting the ILO work programme, the taxonomy structure comprises 400 labour-related terms arranged around 25 subject groups and is used by ILO to describe their webpags content or search the ILO website.

The ILO Thesaurus compiles more than 4,000 concepts around the working world in English, French and Spanish, and grouped under 19 subject categories or facets related to labour issues.

The ILO Taxonomy and Thesaurus are currently managed through Multites, although a new system (implementation: end of 2019) will accept formats as SKOS, SKOS-XL, RDF, OWL, etc. facilitating knowledge sharing within and outside the Organization.

Used by ILO staff, the ILO Taxonomy and Thesaurus are open to the general public, including technical experts, lawyers, academics or researchers.

B. Proposed project and methodology

To allow accessing more content, KOS should be increasingly interrelated.

KOS as terminology assets refer to terms and unique concepts. This is e.g. illustrated in the ISO 25964-2:2011 standard as a "unit of thought" and in the IATE Handbook as follows:

One concept, one entry

- Every entry should deal with a **single concept** (see Annex I for a definition of 'concept'), and all data relating to a given concept should be consolidated in one entry.

Concept (FR: notion): unit of thought constituted through abstraction on the basis of properties common to a set of objects
***N.B.**: Concepts are not bound to particular languages. They are, however, influenced by the social or cultural background.*

Additionally, both kinds of vocabularies include synonyms to disambiguate concepts and relationships between terms, a very common practice for KOS (equivalence / hierarchical / associative relationships) but less in terminology assets.

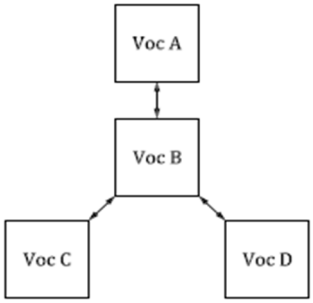
Understanding this, our project focuses first on identifying mappings between concepts / terms / category labels of each of the vocabularies presented above.

Compared to EuroVoc, the very specialised ILO vocabularies present a higher level of granularity. A quite high number of mappings and various content enrichments of each vocabulary are expected to lead to an improved interoperability.

Furthermore, remembering the use of EuroVoc descriptors in IATE, semantic web technologies could help turn traditional terminology assets into linked (open) data contributing to structure some huge bulk of unstructured data available on the web and to implement new features in tools used (or soon to be) by translators.

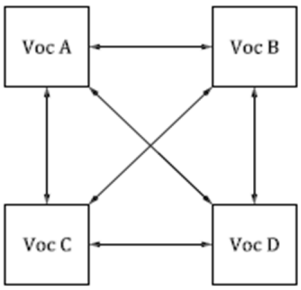
1. An initial project in two phases

- Phase 1, automatic (lexical) alignment to align:
    - All ILO vocabularies with EuroVoc, used as a "hub", as described in the ISO 25964-2:2013 standard (under 6.4).



As such, EuroVoc is used to search in contents indexed with any of the ILO vocabularies; the alignment with a terminology asset is meant to assess consistency and sustainability of use of the contained terms and of the concepts and category labels the other two ILO vocabularies include,

    - Both ILO vocabularies with each other, even if they differ in scope, language and structure; direct mappings must be defined between concepts / terms / category labels of each vocabulary and those of each other vocabulary, as described in the ISO 25964-2:2013 standard:



Aligning with a terminology asset has the same purpose as above.

- Phase 2, automatic (semantic) alignment (same vocabularies).

Additional benefits:

- Streamlining efforts to ensure:
  - Consistency,
  - Understanding inside and between organisations,
- Enriching vocabulary content (see above §IV.B.),
- Inserting mappings,
- Structuring terminology asset(s),
- Implementing these enriched vocabularies and mappings (improving websites search features).

Expected deliverables:

- One or multiple alignment files (in SKOS or EDOAL formats),
- One or multiple files containing evaluation samples,
- A report describing preliminary dataset assessment, designed process and parameters, output alignment files and a final reference point.

All following stages but "g. Alignment design" are valid for both phases of the project.

2. Preliminary assessment

Upon receiving the selected KOS, their initial state is assessed and documented to define whether they are suitable input for the alignment software. Attention is given to both technical and content quality, available languages, presence of duplicates, encoding, estimated pre-processing operations and other aspects. Intended operations and expected final outcomes are listed.

3. Pre-processing

OP's technical team will convert the format of all vocabularies provided by ILO. Input datasets are cleaned up, normalised and transformed to be suitable for automatic alignment.

4. Alignment design

Predefined parameters:

- Main inputs: a pair of datasets or in case of batch alignments many-to-one or one-to-many (don't do many to many),
- Main outputs (in SKOS or EDOAL formats),
- Exact matches only based on perfect equality operator (expected one output) OR
- Close matches based on a designed comparison operator (expected multiple outputs, one per degree of confidence: high, medium, low),
- Comparison operator(s) design (encoded in SILK workbench as a Linking Task).

Alignment comparison operators are of linguistic nature (concepts, language, word, spacing, sequencing, capitalisation, script, encoding, transliteration and others must be considered).

Designing the alignment procedure, all relevant factors from the systematisation presented below are considered.

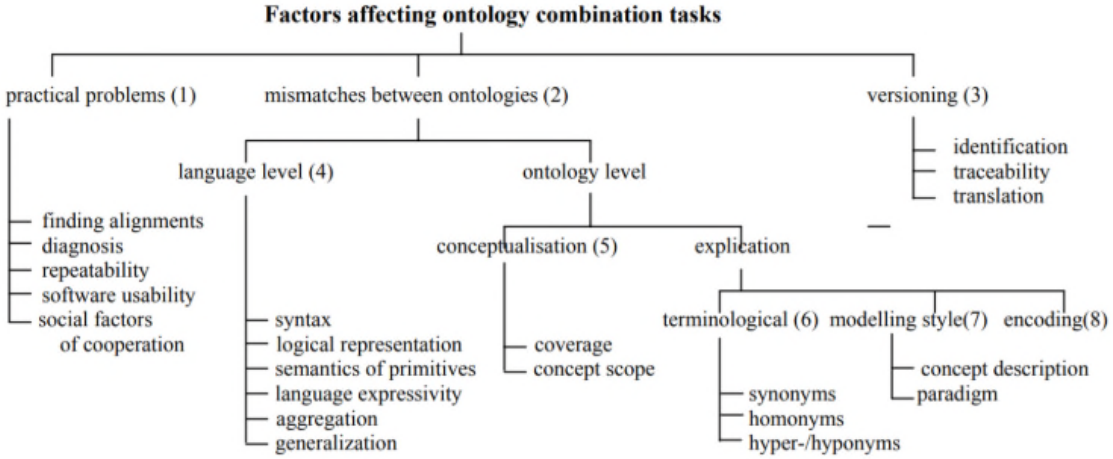


Figure 2: Aspects of ontology integration.

5. Alignment output assessment

The technical team manually assesses the output. All KOS managers will then contribute to the final complete checking and validation of suggested mappings.

6. Final assessment of the exercise

Achieved results are globally evaluated, while the output suitability for publication on the EU Vocabularies website is assessed. The description is supported by statistics generated by extending the code of the *RDF fingerprinter* software to cover the lexicalisation fingerprinting.

Information needed:

- Number of input resources in each dataset and of detected pairs,
- Confidence per dataset and pair,
- Lexicalisation coverage per language in each dataset,
- Relevant dataset fingerprint (covering target lexical resources),
- Number (and percentage) of concept representations included in each mapped source and target dataset.

Further to initial assessments, the report describes operations processed as expected and faced issues. KOS contents are reassessed enlightened by automatic processing results emphasising the lexicalisation and structure qualities, completeness, consistency, etc.

Expected recommendations should focus on:

- Improving KOS,
- Enhancing and automating mapping process,
- Increasing alignment pair numbers and their confidence,
- Interpreting output,
- Further processing steps on the output (automations?),
- Other KOS relevant to this project.

C. Additional tools used

We use(d) the following tools for pre-processing, maintaining, identifying and validating mappings:

- SILK workbench
- VocBench3 (see above §IV.A.1.d.)
- KNIME solutions
- LinkedPipes ETL
- SKOS Play from Sparna
- OpenRefine
- Customised Python scripts and homemade tools.

D. Results

1. Vocabularies considered

	EuroVoc	ILO "gig" glossary	ILO Taxonomy	ILO Thesaurus
Version	4.9.1	(06/2019)	(06/2019)	(06/2019)
Available languages	All EU official languages	EN, ES, FR	EN, ES, FR	EN, ES, FR
Received presentation format	RDF/XML	Excel	XML	HTML
Presentation format converted into	-	RDF/XML	RDF/XML	RDF/XML
Model / Lexicalisation model	SKOS / SKOSXL	SKOS / SKOSXL	SKOS / SKOSXL	SKOS / SKOSXL
Axioms / Number of Triples	3,904,783	716	3,787	6,137
Logical axiom count	1,309,331	111	474	4,833

2. Baseline observations

The original Excel file (glossary) has many imprecise entries impacting the final SKOS version. For instance:

- part-time work / employment (synonym included in label instead of prefLabel/altLabel),
- work-on-demand / on-demand work (two terms for the same concept should be divided in two labels, one defined as preferred),
- contrat permanent (à préférer) | contrat à durée indéterminée (à éviter) (prefLabel vs altLabel?),
- "micro jobbing" (unnecessary quotation marks),

- Capitalisation is inconsistent throughout the original file, especially in Spanish, however with no alignment impact.

3. Results of the process

	Eurovoc	ILO Taxonomy	ILO Thesaurus
ILO Glossary - Equality	X	X	X
ILO Taxonomy - Equality	X	-	X
ILO Thesaurus - Equality	X	-	-
ILO Glossary - Jaccard	-	X	-

4. Preliminary assessment

a) Characterising the input data

ILO files were converted into RDF/XML. All contain labels in English, French and Spanish.

Only the issues mentioned above (see §V.D.b.) could be identified.

Each ILO vocabulary contains one scheme (3 for the thesaurus) covering the complete list of concepts.

All vocabularies were validated in both Prestige and VocBench and no major flaw able to impede the alignment process was identified.

AltLabels are associated to less than a third of concepts in ILO vocabularies with more Spanish ones in the Glossary and none used in the Taxonomy. This inconsistent altLabels coverage makes those irrelevant for alignment.

Even if numerous, ILO Thesaurus scope notes were evaluated as irrelevant for the alignment.

b) Detected issues

In the thesaurus, the use of lower/upper case characters is inconsistent which can be overcome through the standard alignment process with Silk.

In the glossary, several inconsistencies were observed as detailed below:

- The prefLabels are mostly present in all 3 languages but not always,
- The details level presented in each language version can widely vary:
    - The use of notes is not usual here (several prefLabels contain notes),
    - Some labels are too detailed for some concepts.

As such inconsistencies might alter the alignment process, it was decided to align all available languages in parallel to counteract this effect as much as possible.

5. Review of the work performed

The files transformed by the OP were imported into the triple store without additional pre-processing.

a) Aligning ILO vocabularies with EuroVoc

Silk Workbench was used for the alignment with following configuration:

- Paths to source and target set to *skos:prefLabel*
- Transformation: Lower case
- Comparator: Equality
- Link type to *skos:exactMatch*

Equality matching for all ILO vocabularies:

	ILO glossary	ILO taxonomy	ILO thesaurus
Number of source entities	7,243	7,243	7,243
Number of target entities	110	473	4,830
Number of links	7	202	2,067
Valid links	7*	201*	2,030*

*All links have been manually validated as correct (except for deviations induced by concept label definition inconsistencies either in ILO or EuroVoc).

b) Alignments between ILO vocabularies

The exercise aimed to establish to which extent ILO vocabularies share concepts.

The same Silk Workbench approach was used for alignment.

Equality matching for all ILO controlled vocabularies:

Source	Target	Source entities	Target entities	Links	Valid links
ILO thesaurus	**ILO glossary**	4,830	110	9	9
ILO thesaurus	**ILO taxonomy**	4,830	473	523	523
ILO taxonomy	**ILO glossary**	473	110	5	5

*All links have been manually validated as correct (except for deviations induced by concept label definition inconsistencies).

** To ensure alignment's full coverage, the links targeting both broad and narrow concepts having the same labels have been preserved as valid.

Hence, we can conclude the following:

- The taxonomy is well integrated in the thesaurus,
- The glossary is less directly integrated, mainly because the concept of "gig" economy is quite new and constantly evolving (concepts and terms are voluntarily left aside until an official decision).

To identify the reason behind the distance between the Glossary and the Thesaurus an alternative exercise was performed using a token-based approach, the **Jaccard algorithm**:

Source	Target	Source entities	Target entities	Links	Valid links
ILO thesaurus	**ILO glossary**	4,830	110	204	16

The results show an increase of more than 50% of generated links.

6. Preliminary conclusions and possible further steps at this stage

Considering the issues identified (see above):

- Some corrections can be handled, in particular for the ILO Glossary,
- After assessing actual use of both ILO Thesaurus and Glossary, some terms of the Glossary could be added in the Thesaurus.

Validating manually the alignment of EuroVoc with the ILO Thesaurus allowed to evaluate EuroVoc in a different context. Some issues were reported to the team managing EuroVoc, mainly linked with label translation and possible misinterpretations of concepts during translation process.

At this stage of Phase 1, most assumptions seem to be confirmed:

- Phase I (first part):
    - Relevance of mapping EuroVoc and ILO Thesaurus concepts and with ILO Taxonomy category labels (mapping numbers increased) and vice-versa (making ILO Taxonomy and Thesaurus linked data after their SKOS-ification),
    - Need to validate the suggested mappings by all concerned KOS managers,
    - Publication of the mappings between concepts included in EuroVoc and respectively in the ILO vocabularies,
    - EuroVoc linguistic content improvement,
- Phase I (second part):
    - Enhancing the ILO gig glossary, Taxonomy and Thesaurus,
    - Need to:
        - Develop internal synergies between the units managing the vocabularies,
        - Use semantic web technologies for data modelling and the provision of KOS,
        - Make ILO vocabularies linked (open?) data.

6. Broadening the picture

A. Complementing the current results

Phase 1 must be finalised (report drafting and related steps, see above).

Further improving the quality of all selected vocabularies, based on observations and results achieved during the project Phase 1, and ahead as in IATE, the project is defined to contribute to searching enhancements and to improve access to data (discoverability, retrieval, reuse).

Additionally, we see various ways to enlarge the initial scope.

Once Phase 1 is completed, Phase 2 will help improve all considered assets, implement linked data and mappings, and possibly disseminate linked open data. Ultimately, it would help enrich and better structure already used terminology assets.

ILO implementing a new tool to manage KOS might lead to define a third phase or, at least, extend Phase 2.

Moving to linked (open) data and their implementation should foster synergies between (and beyond) organisations and help streamline efforts and assets in both communities of professionals. Ultimately, it would enable them to give their support towards an improved semantic web and contribute to render available increased volumes of valuable information and inferred data that we need as users, decision-makers, linguists, researchers, etc.

Structuring unstructured contents could be an additional phase.

B. A bigger picture for the language professionals?

While publishers disseminate contents with identifiers (URIs, D.O.I., etc.) and insert metadata at the document level to enhance their discoverability, OP inserts metadata at the paragraph level[14]. Similarly, such insertion at sentence/segment level would support the reshaping of linguists' tools. Instead of assembling voluminous relational databases to store contents, unstructured contents disseminated online would benefit from integrating metadata and semantics, and from semantic web technologies.

The main purpose of semantics being to empower extracting knowledge from enormous sets of raw data in various formats, it permits to gain faster and more cost-effective access to meaningful and accurate data, to analyse it and turn it into knowledge.

Additionally, as in IATE, EuroVoc is used for disambiguation purposes in some projects about machine translation (e.g. *Terminology for Machine and Human Translation* and *Machine Translation of Domain-Specific Expressions within Ontologies and Documents*), and to overcome language barriers.

IATE is currently the biggest terminology asset using metadata to structure its content[15]. Streamlining efforts and using URIs[16] instead of labels, would help enrich assets and improve the overall return on investment, contributing towards a further extension of the web of data and KOS enrichment.

Further enabling semantic and terminological interoperability, KOS and terminology assets find some new *raison d'être*:

- Supporting content disambiguation to improve and ease a larger implementation of technologies (e.g. automated tagging and machine translation),
- Benefiting from semantic web technologies (increased discoverability and reuse) and enriching KOS.

Enhancing bridges between terminology assets and KOS, between KOS and corpora[17], and using standard formats, we can foresee in the short-term new tools assembling assets for linguists and drafters, using semantic web technologies and reusing online contents in a click.

In short, further develop both KOS and terminologies, and interlink KOS (also with terminology assets) would be prerequisites.

Acknowledgements

We would like to thank Ms. Christie Damnet (proofreading) for her patience and Mr. Eugeniu Costeţchi (project methodology) and Mr. Mihai Paunescu (KOS analysis and alignments) for all the knowledge passed on.

References

Arcan, Michel. *Machine translations of domain-specific expressions within ontologies and documents.* Insight Centre for Data Analytics.National University of Ireland, Galway. August 2017.

Dechandon, D., Costeţchi, E., Gerencsér, A., Waniart, A "When Terminology Work and Semantic Web Meet". Translating and the Computer 40 Conference, London, 15-16 November 2018.

[14] https://github.com/OpenGovLD/LawAnnotations/wiki/ELI-guides:-reviews-and-excerpts

[15] Denis Dechandon, Eugeniu Costeţchi, Anikó Gerencsér, Anne Waniart, "When Terminology Work and Semantic Web Meet" (2018)

[16] https://www.w3.org/wiki/URI

[17] https://eur-lex.europa.eu/browse/eurovoc.html

Frisendal, T. “State of the Art of Data Modeling?”. Dataversity. June 12, 2017. Accessed 15 October 2019. https://www.dataversity.net/state-art-data-modeling/

Hedden, H. “Taxonomies as Knowledge Organization Systems”. Hedden Information Management. March 17, 2017. Accessed 12 October 2019. https://www.hedden-information.com/taxonomies-as-knowledge-organization-systems/

Hedden, H. “Knowledge Graphs and Ontologies”. Hedden Information Management. May 30, 2019. Accessed 12 October 2019. https://www.hedden-information.com/knowledge-graphs-and-ontologies/

IATE Handbook taskforce. IATE Handbook. https://iate.cdt.europa.eu/iatenew/handbook.pdf. Accessed 15 October 2019.

ISO 25964-1:2011, Information and documentation — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval. https://www.iso.org/standard/53657.html

ISO 25964-2:2013, Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies. https://www.iso.org/standard/53658.html

Kockaert, Hendrik J., Steurs, Frieda. *Handbook of Terminology. Terminology tools.* Frieda Steurs, Ken de Wachter, Evy de Malsche. John Benjamins Publishing Company, Amsterdam, 2015, p. 224.

Kuckartz, A. “A technical implementation guide (2n edition)”. ELI guides: reviews and excerpts. October 2, 2018. Accessed 15 October 2019. https://github.com/OpenGovLD/LawAnnotations/wiki/ELI-guides:-reviews-and-excerpts

Martín-Chozas, P., Rodríguez-Doncel, V. Towards a Linked Open Data Cloud of Language Resources in the Legal Domain. Law via the Internet Conference, Florence, 11-12 October 2018.

Vasiljevs, A., Pinnis, M. Terminology for machine and Human Translation. April 2019.

W3C. Semantic Web. https://www.w3.org/2001/sw/wiki/Main_Page

Zeng, Marcia. (2008). Knowledge Organization Systems (KOS). Knowledge Organization. 35. 160-182. 10.5771/0943-7444-2008-2-3-160.