Umanistica Digitale – ISSN: 2532-8816 – n. 18, 2024
M. Ramos, R. Costa – *Applying Text Mining Methods to Construct a Domain Ontology from Definitions*
DOI: http://doi.org/10.6092/issn.2532-8816/19284

Applying Text Mining Methods to Construct a Domain Ontology from Definitions

Margarida Ramos

NOVA CLUNL - Centro de Linguística da Universidade NOVA de Lisboa, Lisboa, Portugal
mvramos@fcsh.unl.pt

Rute Costa

NOVA CLUNL - Centro de Linguística da Universidade NOVA de Lisboa, Lisboa, Portugal
rute.costa@fcsh.unl.pt

Abstract

This paper aims to describe a text-mining approach on a domain corpus (cork) within the theoretical framework of the dual dimension of terminology to create a terminological dictionary and correlate it with an ontology. We will make some considerations on (i) domain specificities; (ii) lexical markers; (iii) automatic corpus processing using Sketch Engine; (iv) representation of lexical networks using CmapTools; and (v) representation of the concept system using Protégé. The goal of the ontology is to logically support the coherence and quality of the natural language definitions contained in the terminological resource.

Keywords: terminology; definition; domain-ontology; knowledge-rich information; domain corpus; terminological dictionary.

Questo articolo si propone di descrivere un approccio di text-mining su un corpus di dominio (sughero) nel quadro teorico che contempla la dimensione duale della terminologia al fine di sviluppare un dizionario terminologico e correlarlo ad un'ontologia. Verranno fatte alcune considerazioni su (i) specificità del dominio; (ii) marcatori lessicali; (iii) trattamento automatico del corpus utilizzando Sketch Engine; (iv) rappresentazione delle reti lessicali tramite l'utilizzo di CmapTools; e (v) rappresentazione del sistema concettuale con Protégé. L'obiettivo dell'ontologia è quello di supportare, dal punto di vista logico, la coerenza e la qualità delle definizioni in linguaggio naturale contenute nella risorsa terminologica.

Parole chiave: terminologia; definizione; ontologia di dominio; informazione ricca di conoscenza; corpus di dominio; dizionario terminologico.

Introduction

This paper aims to demonstrate a method for developing a terminological dictionary rooted in a domain ontology. We will describe the methods employed on a domain-specific corpus to capture specialised lexical and conceptual knowledge, exploiting this information to develop a dedicated ontology. The terminological resource will offer a linguistic description of the specialised concepts, relying on the formal definitions of the concepts that compose the cork ontology, OntoCork [37].

The foundation for constructing an ontology to represent the knowledge of the cork transformation sector is based on two main perspectives: (i) introducing an innovative dimension to the domain, and (ii) creating a tool tailored for experts, future experts, and linguists. This tool aims to provide a shared understanding of the specific domain, which can serve as a unifying framework to address issues such as communication challenges [44].

Our focus will be on the cork industry. Our work revolves around specialised activities divided into four sub-sectors, each focused on different cork-related tasks: (1) the preparation of cork; (2) the transformation of cork; (3) the granulation and (4) the agglomeration of cork products. The texts being analysed report on these activities and constitute a subcorpus of CorkCorpus, a comprehensive domain corpus designed from scratch. By analysing the corpus, we identified several nuclear concepts of the domain that are either not defined or only partially defined in normative documents, despite their terms being widely used by domain experts [14][20]. This fact derives from the intersubjectivity shared by a given community of expertise; such knowledge does not need to be defined, as it is commonly understood among the experts. Therefore, organising knowledge becomes challenging when the terminologist-linguist lacks expertise in the specific domain being studied. One potential approach to address the missing information is to develop strategies for exploring the corpus to capture definitional contexts. For analysing these definitional contexts, we employed natural language processing methods focused on text analysis, which is a key task within text mining [39].

In our text analysis, we used Sketch Engine to observe lexical-semantic markers—specifically, co-occurring lexical units frequently found in contextual definitions. These patterns serve as entry points in our study to infer lexical-semantic relationships between terms, thereby offering data to interpret the expert knowledge conveyed in texts.

The methodology outlined in the present paper relies on the semantic labels observed in the corpus, which serve as the central focus for constructing the domain ontology. From our viewpoint, this approach embodies the double dimension of Terminology, where linguistic and conceptual information complement each other without overlapping. Accordingly, the essential initial step in understanding expert conceptualisations involves interpreting texts generated within specialised communication contexts, as texts serve as the primary means for knowledge transfer. Hence, the relevance of conducting linguistic analysis before proceeding with conceptual organisation is emphasised.

1. Description of the activities involved in the domain of cork

The cork industry encompasses four primary sub-sectors: preparation, transformation, granulation, and agglomeration of cork products.

1. Preparation: This stage involves several steps, including 'slicing' (*traçamento*), 'stacking' (*empilhamento*), 'boiling' (*cozedura*), and 'stabilizing' (*estabilização*) the cork [14][6].
2. Transformation: This sub-sector focuses on manufacturing natural cork stoppers and producing cylindrical batons from granulated cork, which are then used to fabricate agglomerated cork stoppers and components for technical stoppers. These products often undergo 'finishing processes' (*acabamento*) [6][20].
3. Granulation: This activity includes breaking down cork into granules, which are used in various applications.
4. Agglomeration: This involves combining granulated cork into products used in the construction industry, as well as in the automobile and aeronautical sectors, among others[19][13][15].

We will focus on the transformation sub-sector, as our study centres on the production of natural cork stoppers. The journey from cork extraction to the final product involves several stages, which vary based on the type of stopper being produced. The overall production process of cork stoppers is divided into three main stages: 'debarking' (*descortiçamento*), 'manufacturing the stopper' (*fabrico da rolha*), and 'finishing the stopper' (*acabamento da rolha*). Each of these stages encompasses various specific processes[14][20][5].

The stages of debarking and finishing the stopper are identical for both natural and agglomerated cork stoppers. The manufacturing of natural cork stoppers pertains exclusively to the transformation activity, while agglomerated cork stoppers involve additional fabrication processes within the granulation and agglomeration activities. These latter processes will not be covered in this study.

The production of natural cork stoppers is a key transformation activity. These stoppers are obtained from thick, rectangular pieces of cork known as 'stripes' (*rabanadas*) through a process called 'punching' (*brocagem*). Experts explain that to achieve the rectangular shape, cork planks must first be 'sliced' (*rabaneadas*) before the stripes are punched out [14][20]. At this stage, immediately after being punched from the stripe, the stopper is a semi-manufactured product, still far from being a finished item.

A semi-manufactured natural cork stopper undergoes several additional operations to become a finished product. This is where the finishing process plays a crucial role in the transformation activity. Cork stoppers can be sold in either a semi-finished or fully-finished state. Clients, such as wineries, may purchase them unfinished or ready for use, depending on their specific needs and capabilities to complete the finishing process [12]. In summary, a semi-finished stopper is a stopper that has undergone one or various 'finishing treatments' (*tratamento de acabamento*) during the 'finishing process' (*processo de acabamento*). These treatments may include 'rectifying' (*rectificação*), 'washing' (*lavação*) followed by 'drying' (*secagem*), or even 'sealing' (*colmatagem*), but exclude the 'final treatment' (*tratamento final*). At this stage, the semi-finished stopper can either be sold, packed, and transported, or it can continue through the finishing process until it is ready for use. To be considered a finished product, the stopper must undergo final treatments such as 'branding' (*marcação*) and/or surface coating with silicone or paraffin wax [19][9][13][30].

2. The Cork corpus: purpose, criteria, and description

In our study, one of the primary objectives is to discern the meaning of lexical units through their contextual usage. However, it is important to note that many of these units are not solely lexical but are terms, each designating a specific concept. As terminologists, our pursuit focuses on identifying designations that hold a mono-referential meaning—a scenario where one term corresponds to one concept.

The corpus comprises texts produced by experts, namely technical, and scientific professionals in the cork industry. The collection of texts was based on strict criteria specific to terminological work [36], highlighting the specialised context of text production as a fundamental element. Considering the idea that 'corpus size depends on the research question and the kind of linguistic features we want to investigate' [8], we assume that the size of our domain-specific corpus (see Figure 1 and Table 2) adequately fulfils the objectives of this study as it remains representative of the population it draws from [10]. In essence, our emphasis on selectivity in text collection prioritises quality over quantity, aligning with the principles advocated by the Corpus Linguistics community [43][24][45] and domain-specific corpora builders[7][1][4].

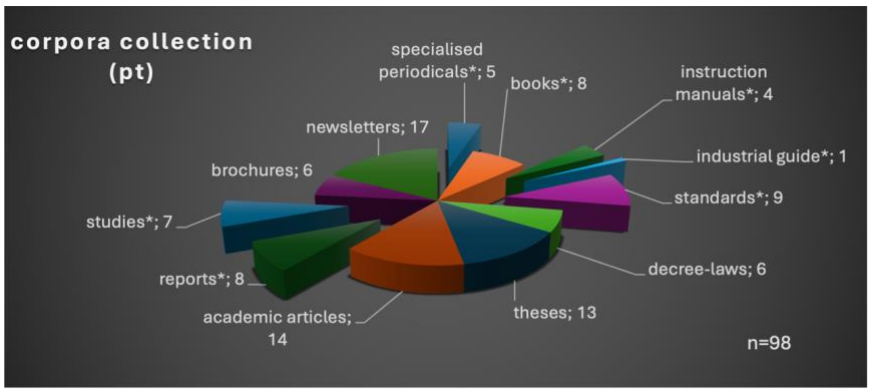

Figure 1: Corpora collection.

The quantitative data derived from the 98 documents of the corpus is presented in Table 1.

Data	Total number
Tokens	1,706,947
Types	91,250
Forms (sequences of characters)	1,213,625
Sentences	44,683

Table 1. Quantitative data of the corpus.

The Cork Corpus was compiled by collecting texts produced within the cork industry. The purpose of creating a domain-specific corpus is to linguistically analyse the discourse of experts in order to extract information that may represent expert conceptualisations beyond verbal expression. The *internal* and *external* criteria [3][32] used to build the Cork Corpus are systematised in Table 2.

Criteria	Purpose/description
Degree of specialisation	Produced by experts and semi-experts
Source validation	Entities recognised as an authority
Type	**Technical explanatory; normative**
Content adequacy	On cork/Cork stoppers
Synchronism (≤ 10 years)	Given the fast evolution of technology

Table 2. Internal and external criteria of the Cork Corpus.

Among the criteria outlined in Table 2, the foremost consideration for compiling the corpus to facilitate our analysis was the communicative context in which the texts were produced. Our primary focus in the linguistic analysis is to observe texts typically comprising technical-explanatory content. This includes normative texts, emphasised in bold within Table 2, alongside content tailored for economic and financial spheres. Notably, the latter texts are authored by domain experts specifically for governmental institutions' experts. We opted for the latter setting due to the inclusion of glossaries and definitions authored by experts, ensuring an *a priori* validation of the extracted terminological data.

The Cork Corpus consists of 98 texts written in European Portuguese. These publicly available texts are authored not only by experts from various organisations within diverse domains associated with the cork industry but also from different sectors, including scientific, industrial, techno-professional, certifying, regulating, and commercial fields (see Figure 1).

Adhering to the mentioned criteria, we curated a *balanced corpus* [4] encompassing a wide range of domain-specific texts—, i.e. a collection of texts ranging from highly technical content aimed at experts to more accessible texts tailored for non-experts despite the consistent use of domain-specific terminology in the latter. The rationale behind selecting texts aimed at non-experts is tied to their richness in definitional contexts or contexts pointing towards concepts, given the different knowledge degrees held by producers and recipients.

During the corpus exploration and linguistic data analysis, our main focus was on 44 texts generated within two distinct communicative settings, where definitional contexts are commonly found: (i) domain expert to other domain experts and (ii) expert to semi- experts. The more significant the knowledge gap between the author-expert and his/her audience, the more definitions and contextual definitions tend to be produced. The remaining 54 texts served as a 'reference corpus' [4], enabling a comparative analysis of the terminological data extracted from the subcorpus comprised of 44 texts (=1,022,570 tokens).

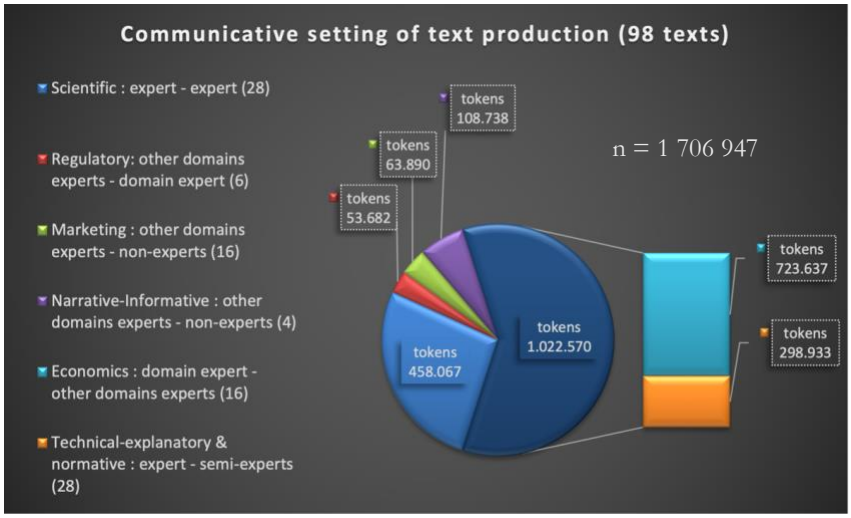

Figure 2. Subcorpus under focus (44 texts=1,022,570 tokens) based on two communicative settings of text production: (1) Economics, and (2) Technical-explanatory & normative.

As depicted in Figure 2, the subcorpus on which we have conducted our main terminological analysis unfolds in 16 texts produced in the Economics setting (domain expert – other domains experts), with 298,933 tokens, and 28 texts in the Technical-explanatory & normative (expert – semi-experts), with 723,637 tokens. Although classified under two (2) communication settings here, the technical content of the subcorpus is well-balanced, as illustrated in Figure 1 with an asterisk.

3. Methods

The methodology employed in this study is corpus driven. To analyse the corpus, we employed Sketch Engine[1] to identify and systematise lexical-semantic relations. During the corpus analysis, two types of valuable knowledge-rich information [28] emerged: (1) definitions and (2) definitional contexts. According to [21], 'definition' is the representation of a concept by an expression that describes it and differentiates it from other related concepts. Definitions are one of the components of the glossaries' microstructure, typically situated at the conclusion of normative texts, aiming to establish consensus within the cork community. Definitional contexts within texts also contain rich and essential information—terms, collocations, and lexical markers—that are crucial for understanding specific concepts.

Our method involves two primary stages:

[1] https://www.sketchengine.eu/ [Accessed 2024-10-16].

i. Based on linguistic analysis of lexical markers and their corresponding lexical-semantic relations observed among co-occurring terms along the syntagmatic axis, we organise the results into lexical maps using CmapTools. [2]

ii. Building upon the previous stage, we progress to conceptual analysis and subsequent formal representation. The conceptual analysis forms the basis for identifying conceptual relations derived from interpreting the lexical-semantic relations observed between two terms. Employing deductive mechanisms based on the Aristotelian formula 'X=Y+DC', we infer conceptual relations, such as associative types, and identify characteristics that facilitate the process of creating concept systems to develop OntoCork [37].

3.1 Terminological data extraction

We processed the corpus using Sketch Engine, employing it to compile, annotate, and execute advanced searches using the Corpus Query Language[3] (CQL) syntax, which incorporates the application of regular expressions (regexes).

The most frequent terms in the Cork Corpus are '*cortiça*' (cork; F=16,127), '*rolha*' (stopper; F=5,862) and '*rolhas*' (stoppers; F=1,221). Given their significant frequencies, we analysed the contexts of their occurrence within the subcorpus of 44 texts. Initially, we used the Word Sketch[4] tool, leading us to the identification of candidate terms such as '*ROLHA COLMATADA*' (colmated stopper) denoted in capital letters. Subsequently, we conducted simple queries (concordances) aimed at locating polylexical terms that incorporate adjectives, for which the tool was parametrised to capture linguistic forms labelled as adjectives, as follows:

```
Concordance: "advanced"
        Part-of-speech: "any"
        Query type: "word"
        Word: "rolha"
Filter context: "Part-of-speech context"
        Only keep lines with
        "all" of "adjective" within "1" Tokens "right".
```

This focus on adjectives ties with the nature of a '*rolha*' (stopper) being a manufactured object, thereby various types and manufacturing states are expressed in texts concerning this object. Commonly, these types and states are conveyed by means of qualities or attributes expressed through adjectives, due to their function as noun modifiers. Table 3 displays the results of the most frequent adjective forms (among the top 300), a few of which we hypothesise being constituents of polylexical terms, such as *'natural'* (natural; F=1,944) and *'técnico'* (technical; F=510), due to their highest frequency in the grammatical category we have focused on in the Word list[5] tool.

[2] https://cmap.ihmc.us/ [Accessed 2024-10-16].

[3] https://www.sketchengine.eu/documentation/corpus-querying/ [Accessed 2024-10-16].

[4] https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/ [Accessed 2024-10-16].

[5] https://www.sketchengine.eu/guide/glossary/?letter=W [Accessed 2024-10-16].

Adjective form (pt)	Adjective form (en)[6]	Absolute Frequency	Frequency per million
natural	natural	1,944	1,135.08
técnico	technical	510	297.78
seco	dried	434	253.40
cilíndrico	cylindrical	207	120.86
lenticular	lentiform	161	94.00
suberoso	subereous	160	93.42

Table 3. The most frequent adjective forms (within the first 300 results) that were deemed to correspond to terms (or part of polylexical terms).

After identifying the most frequent morphosyntactic structures of terms through the Word Sketch tool (see Table 4 for the top 5), we decided to improve our search for terms and definitions using advanced queries. Specifically, we employed CQL expressions using regexes (see Section 3.2), aiming to capture knowledge-rich contexts (KRCs) [28], e.g., definitions (found in context) and definitional contexts, i.e., contexts providing explanations of the concept. Such contexts are invaluable for the comprehension and/or elaboration of proper definitions.

rolha' + ADJ (pt)	ADJ + stopper (en)	Occurrences in the corpus
rolha natural	natural stopper	114
rolha técnica	technical stopper	63
rolha capsulada	capsulated stopper	54
rolha cilíndrica	cylindric stopper	54
rolha aglomerada	agglomerated stopper	49

Table 4. The morphosyntactic structure of 5 polylexical candidate terms [N + ADJ] identified via the Word Sketch tool.

An additional rationale for capturing KRCs is the prospect of observing linguistic expressions pointing to or relating terms. These contexts hold significant relevance as they provide information aiding in deducting the concepts designated by those terms. They enable terminologists to infer the experts' knowledge by identifying recurring linguistic patterns prevalent in texts. For instance, experts consistently employ specific domain-specific linguistic expressions to relate terms, thereby indirectly signifying the underlying concepts. The two definitions presented in Table 5 serve as illustrations supporting this assertion.

[6] Literal translation.

Definitions extracted from the corpus (pt)	Literal translation in English (en)	
ROLHA LAVADA: Rolha que foi submetida a um tratamento químico com o objectivo de desinfectar e/ou homogeneizar a cor e/ou branquear.	WASHED STOPPER: **Stopper** that was submitted to chemical treatment with the aim of disinfecting and/or homogenising the colour and/or bleaching.	(1)
ROLHA PONÇADA[7]: Rolha cuja superfície lateral foi submetida a uma operação de abrasão para a tornar cilíndrica ou diminuir o seu diâmetro.	SIDE-SURFACE SANDED STOPPER: **Stopper** whose side surface was submitted to an abrasive operation to make it cylindrical or to reduce its diameter.	(2)

Table 5. Linguistic expressions commonly used by experts.

The linguistic patterns underlined in Table 5 represent lexical markers, which are linguistic expressions commonly found within definitions. These lexical markers hold significant importance as they facilitate linguists in deducing specialised knowledge. They indicate lexical-semantic relations between terms, thereby offering valuable data for interpreting the expertise conveyed within texts. For instance, the linguistic expression '*foi submetida a'* (was submitted to) serves as a lexical marker relating the term '*rolha'* (stopper) to two candidate terms: (1) '*tratamento químico*' (chemical treatment) and (2) '*operação de abrasão*' (abrasive operation), providing insights into the treatments or operations associated with the object. Additional insights regarding lexical markers indicating lexical-semantic relations are provided in Section 4.

3.2 Exploring the corpus with text-mining methods

Building upon the identified patterns (Table 5), we conducted further exploration within the subcorpus using either (i) refined CQL expressions previously employed or (ii) newly formulated CQL expressions. It is crucial to note that both approaches stemmed from the outcomes of multiple rounds of advanced queries. This continuous refinement of regexes constituting the CQL expressions led us to develop an iterative text-mining method. This method allowed us to align with recurring patterns commonly found in specialised texts—texts authored by and for domain experts. Consequently, this facilitated the extraction of the terminological data we sought, particularly definitional contexts.

In this paper, we aim to highlight two distinct CQL expressions that have demonstrated effectiveness in identifying lexical-semantic relations (inferred from lexical markers) between terms. Additionally, these expressions have proven valuable in uncovering definitional contexts wherein the generic term is expanded within its syntax (e.g., bolded terms in Table 5)

The first CQL expression has the following structure:

(1) "rolha"[tag="V.P.*SF"]

whose formulation aims to match the pattern: ONLY the linguistic string '*rolha'* (stopper) followed by ANY past participle ONLY in the singular and feminine inflection.

[7] The adjective '*ponçada'* does not have an equivalent in English. Based on the analysis of the characteristics stated in the definition, we propose the adjective phrase *side-surface sanded.*

To craft CQL(1), we focused on recurring linguistic expressions frequently used by experts, particularly instances of the past participle co-occurring with a term (as demonstrated revealed in Table 5). This query matched 69 linguistic strings and highly productive patterns for identifying lexical markers. For instance, it matched expressions like '*x foi submetida a y*' (x was submitted to y) and terms conforming to our search patterns of [Noun + Past Participle], e.g., '*x acabada*' (finished x) or '*x terminada*' (finalised x). Here, '*x*' represents a term, while '*y*' corresponds to a structure that consistently encapsulates knowledge-rich information—insights provided by experts, allowing us to grasp their *conceptualisations* [34].

Based on the satisfactory results of CQL(1), we decided to expand its formulation into CQL(2):

(2) "rolha"[(tag="D.*"|(tag="S.*")]?[tag="A.*"]?"cortiça" ? [ ] {0,4}"rolha" [ ] {0,4} [tag="V.P.*SF"]

In this scenario, our aim is to identify a context where the term '*rolha'* (stopper) might appear alongside any determiner OR any preposition OR NOT, potentially followed by any adjective OR NOT. Subsequently, the term '*cortiça'* (cork) may or may not occur, succeeded followed by none or up to four *tokens*[8] or strings of tokens before matching the pattern outlined in CQL(1).

The rationale behind the pattern(s) possibly matched by CQL(2) is rooted in the structure of *intensional definitions* [21]. These definitions typically start with the reference to the closest generic concept, followed by characteristics that hierarchically or causally differentiate the specific concept being defined, as exemplified by the boldened words in Table 5 and Table 6.

CQL(2)proved highly effective in capturing terminological data. Among the 55 linguistic strings obtained, 44 partially comprised either descriptions or definitions. Among the remaining 11 strings, 7 are repeated definitions as a consequence of the operator '*?*', whose function enables matching both zero and one occurrence of the previous element in the query. The final 4 strings were polylexical terms found in titles.

Out of the set of descriptions and definitions extracted semi-automatically from the corpus, we chose to focus on ten (10) definitions (see Table 6) for both linguistic and conceptual analysis. We chose these definitions according to the following criteria: (1) the broad scope of the generic term, (2) information regarding the object's structure, including shape and components, (3) different types of substances like natural cork or agglomerated cork, and (4) examples of stoppers classified by their finishing process. Moreover, the inclusion of two identical generic terms, "rolha" [stopper], in lines 1 and 2 is justified by the additional information they provide in both definitions, specifically describing the shape of the piece of cork and the knowledge of it being a product.

#	**10 definitions extracted from the Cork Corpus (pt)**	**10 definitions (literal translation from pt)**
1	rolha **Produto** obtido da cortiça natural e / ou de cortiça aglomerada, constituído por uma ou mais peças, destinado a vedar garrafas ou outros recipientes e a preservar o seu conteúdo. (5.1 - NORM)	stopper **Product** obtained from natural cork and / or agglomerated cork, consisting of one or more pieces, intended to seal bottles or other containers and to preserve their contents. (5.1 - NORM)

[8] https://www.sketchengine.eu/my_keywords/token/ [Accessed 2024-10-16].

2	ROLHA **peça de cortiça**, em geral cilíndrica, troncocónica ou prismática quadrangular, por vezes de arestas laterais boleadas ou chanfradas, constituída por um ou vários elementos colados e destinada a vedar os recipientes ou a contribuir para a sua estanquicidade (7.8 – TECH)	STOPPER **piece of cork**, usually cylindrical, conical or prismatic quadrangular, sometimes with rounded or chamfered lateral edges, consisting of one or several glued elements and intended to seal the containers or contribute to their water tightness. (7.8 – TECH)
3	rolha de cortiça natural **Rolha** totalmente constituída por cortiça natural. Nota: As rolhas naturais que tenham sido submetidas à operação de colmatagem (ver 6.5.5) são comummente designadas por rolhas naturais colmatadas. (5.5 – NORM)	natural cork stopper **Stopper** consisting entirely of natural cork Note: Natural cork stoppers that have been submitted to the sealing operation (see 6.5.5) are commonly referred to as colmated natural stoppers. (5.5 – NORM)
4	rolha de cortiça natural colmatada A rolha de cortiça natural colmatada è uma **rolha** feita de cortiça natural em que são obturadas as suas lenticelas com uma mistura de colas e pó de cortiça proveniente dos acabamentos dimensionais das rolhas de cortiça natural. (6.1 – REP)	colmated natural cork stopper The colmated natural cork stopper is a **stopper** made of natural cork in which its lenticels are filled with a mixture of glues and cork powder from the dimensional finishing processes of natural cork stoppers. (6.1 – REP)
5	rolha de cortiça aglomerada **Rolha** obtida pela aglutinação de granulado de cortiça com dimensão compreendida entre 0,25mm e 8mm, com adição de ligantes, através de extrusão ou moldagem e composta, pelo menos, por 51 % de granulado de cortiça, em peso. (5.5 – NORM)	agglomerated cork stopper **Stopper** obtained by the agglutination of cork granules with a size between 0,25 mm and 8 mm, with addition of binders, by means of extrusion or moulding and composed of at least 51% by weight of cork granules. (5.5 – NORM)
6	rolha aglomerada: **peça de cortiça** aglomerada, *obtida por* extrusão ou moldagem (3.1 – STUD)	agglomerated stopper: piece of agglomerated cork, obtained by extrusion or moulding (3.1 – STUD)
7	rolha n+n **Rolha** formada por um corpo de cortiça aglomerada e "n" discos de cortiça natural colados num ou em ambos os topos.	n+n stopper **Stopper** formed by a body of agglomerated cork and "n" disks of natural cork glued to one or both ends.

	Nota: Nesta designação, "n" indica o número de discos utilizados. (5.5 – NORM)	N.B.: In this designation, "n" indicates the number of disks used. (5.5 – NORM)
8	rolha técnica As rolhas técnicas são constituídas por um corpo de cortiça aglomerada, muito denso, com discos de cortiça natural colados no seu topo – ou em ambos os topos. As rolhas técnicas com um disco em cada topo são designadas rolhas técnicas 1+1. Com dois discos de cortiça natural em cada topo chamam-se rolhas técnicas 2+2, e com dois discos em apenas um dos topos chamam-se rolhas técnicas 2+0. (6.1 – REP)	technical stopper Technical stoppers are composed of a very dense body of agglomerated cork with disks of natural cork glued to one end – or to both ends. Technical stoppers with one disk on each end are called 1+1 technical stoppers; those with two disks of natural cork on each end are called 2+2 technical stopper; and those with two disks glued at only one of the ends are called 2+0 technical stoppers. (6.1 – REP)
9	rolha boleada **Rolha** cujas arestas de um ou dois topos foram arredondadas, por abrasão. (5.5 – NORM)	rounded stopper **Stopper** whose edges of one or two ends were rounded by abrasion. (5.5 – NORM)
10	ROLHA MARCADA Rolha cuja superfície lateral ou topos foram marcados a tinta ou a fogo. (7.6 – TECH)	MARKED STOPPER Stopper whose lateral surface or ends were marked in ink or by fire (7.6 – TECH)

Table 6. Ten (10) definitions to organise a typology of cork stoppers.

For the purpose of this paper, we will solely consider one specific definition—namely, that of <Rolha de cortiça natural>[9] (natural cork stopper), registered in Table 6, line 3—to demonstrate our linguistic and conceptual analyses. To enhance clarity, we will conduct the linguistic and conceptual analyses using candidate terms in English, while including the original text in Portuguese. However, visual representations of text interpretation, such as maps, as well as semi-formal and formal descriptions of conceptual information, will be presented exclusively in English.

[9] We use typographic conventions to differentiate the axis of analysis under focus. With this procedure, terms, concepts, characteristics, and conceptual relations are clearly differentiated from each other. Concepts are written in two different ways depending on the axis of analysis: (i) within angle brackets and the first letter capitalised, e.g., <Concept>; <Concept_1> for the linguistic and conceptual analyses, or with CamelBack notation for the ontology representation, e.g., `ConceptExample`; (ii) characteristics are written within forward slashes, e.g., /characteristic/; conceptual relation identifiers are written in italics with an underscore separating each form, e.g., *has_relation*; and (iii) conceptual relations are written with CamelBack notation, e.g., `hasRelation`.

4. The linguistic analysis

	Definition in context
<Rolha de cortiça natural>	Rolha totalmente constituída por cortiça natural. Nota: As rolhas naturais que tenham sido submetidas à operação de colmatagem (ver 6.5.5) são comummente designadas por rolhas naturais colmatadas.
<Natural cork stopper>	stopper consisting entirely of natural cork Note: Natural cork stoppers that have been submitted to the sealing operation (see 6.5.5) are commonly referred to as colmated[10] natural stoppers.[11]

	Analysis	**Lexical marker (LM)**	**Lexical-semantic relations**	**Interpretation**
LINGUISTIC DIMENSION	natural cork stopper [is a] stopper rolha de cortiça natural [é uma] rolha	'is a' = Ø 'é uma' = Ø	**HYPERNYMY - HYPONYMY**	stopper [GENERIC] natural cork stopper [SPECIFIC] rolha [GENERIC] rolha de cortiça natural [SPECIFIC]
	natural cork stopper [consists entirely of] natural cork rolha de cortiça natural [totalmente constituída por] cortiça natural	'consisting entirely of' 'totalmente constituída por'	**HOLONYMY-MERONYMY**	natural cork stopper [OBJECT] natural cork [STUFF] rolha de cortiça natural [OBJECT] cortiça natural [STUFF]
	natural cork stopper [is submitted to] the sealing operation rolha de cortiça natural [é submetida a] operação de colmatagem	'submitted to' 'submetida a'	**HOLONYMY-MERONYMY**	sealing operation [ACTIVITY] ? = [FEATURE] operação de colmatagem [ACTIVITY] ? = [FEATURE]

[10] Despite the inexistence of the adjective 'colmated' in English, we have found the term 'colmated corks' used as an equivalent for *rolhas colmatadas* in texts produced by native English speakers [42].

[11] Literal translation. Source: Cork Corpus, text 5.5 – NORM.

	colmated natural stopper [is a] natural cork stopper rolha natural colmatada [é uma] rolha de cortiça natural	'commonly referred to as' same as = 'is a' 'comummente designada por' same as = 'é uma'	**HYPERNYMY - HYPONYMY**	natural cork stopper [GENERIC] colmated natural stopper [SPECIFIC] rolha de cortiça natural [GENERIC] rolha natural colmatada [SPECIFIC]
	colmated natural stopper [results from] the sealing operation rolha natural colmatada [resulta de] operação de colmatagem	results from = inferred from 'submitted to' resulta de = inferred from 'submetida a'	**HOLONYMY-MERONYMY**	sealing operation [ACTIVITY] colmated = [FEATURE] operação de colmatagem [ACTIVITY] colmatada = [FEATURE]

Table 7. Linguistic analysis of the definition of the concept of <Natural cork stopper>.

With Table 7, we present the initial phase of our study, detailing the deconstruction of the definition and conducting its linguistic analysis. Our objective is to analyse the lexical-semantic relations among terms. The definition of <Natural cork stopper> is presented in the primary sentence, accompanied by supplementary encyclopaedic information in the form of a note. While the first sentence provides essential insights into the composition of a <Natural cork stopper>, the encyclopaedic information explains the nature of the object when submitted to a specific operation.

Through our analysis, the initial inference drawn is that a <Natural cork stopper> 'is a stopper'. Here, 'is a' functions as a lexical marker relating term A (natural cork stopper) and term B (stopper), reflecting a clear hypernym-hyponym relation, wherein 'natural cork stopper' serves as the hyponym of the hypernym 'stopper'. This analysis can be represented as follows:

(1) [natural cork stopper] HYPONYM / SPECIFIC is a [stopper] HYPERNYM / GENERIC

[1] [*rolha de cortiça natural*] HYPONYM / SPECIFIC *é uma* [*rolha*] HYPERNYM / GENERIC

The next insight derived from our linguistic analysis stems from the expression 'stopper consisting entirely of natural cork' (Table 7, line 2). Within this context, the lexical marker 'consisting entirely of' (*totalmente constituída por)* establishes a relation between the terms 'natural cork' and 'stopper'. While 'natural cork' defines the substance of the object, 'stopper' refers to the object itself. The relation between the terms is undeniably meronymy, falling under the subtype [OBJECT-STUFF] [25], and can be represented as follows:

(2) [stopper] OBJECT consisting entirely of [natural cork] STUFF

[2] [*rolha*] OBJECT *totalmente constituída por* [*cortiça natural*] STUFF

Expanding upon representation (2), e can delve into another interpretation. Given that 'stopper' serves as the generic term encompassing 'natural cork stopper'—a deduction drawn from the 'is a' lexical marker—, we can reformulate this information into the following representation:

(3) [natural cork stopper] OBJECT consisting entirely of [natural cork] STUFF

[3] [*rolha de cortiça natural*] OBJECT *totalmente constituída por* [*cortiça natural*] STUFF

The second sentence, presented as a note within the definition, provides additional information obtained by analysing the statement 'natural cork stoppers that have been submitted to the sealing operation.' In this instance, the lexical marker is 'submitted to' (*submetidas à*), establishing a relationship between the term 'natural cork stopper' (*rolha natural)* and the term 'sealing operation' (*operação de colmatagem)*. 'Sealing operation,' denoting a specific activity, is related by the lexical marker to 'natural cork stopper,' an object we have already identified (see Map 1 – Figure 3). Through the interpretation of their meanings, we can deduce that the lexical-semantic relation established here is meronymy, categorised as subtype [ACTIVITY-FEATURE] (see Map 1.1 – Figure 4). We concluded that the lexical marker 'are referred to as' expresses the lexical-semantic relation of hypernymy-hyponymy. This determination was based on two factors: (i) the terms related by this lexical marker are 'natural cork stopper' and 'colmated natural stopper,' and (ii) the meaning of the former encompasses a broader scope than that of the latter. This final interpretation is represented as follows:

(4) [colmated natural stopper] HYPONYMY/SPECIFIC is a [natural cork stopper] HYPERNYMY/GENERIC

[4] [*rolha natural colmatada*] HYPONYMY/SPECIFIC *é uma* [*rolha de cortiça natural*] HYPERNYMY/GENERIC

Consistently, we have employed the same methodology for linguistic analysis, particularly in identifying lexical markers denoting lexical-semantic relations between two terms. Determining the exact lexical-semantic relation expressed by a given lexical marker is not always straightforward, especially when identifying meronymy subtypes. In such cases, the interpretation of each meaning of the related pair of terms becomes crucial. Understanding each of these meanings is pivotal in delineating the specific subtypes of meronymy.

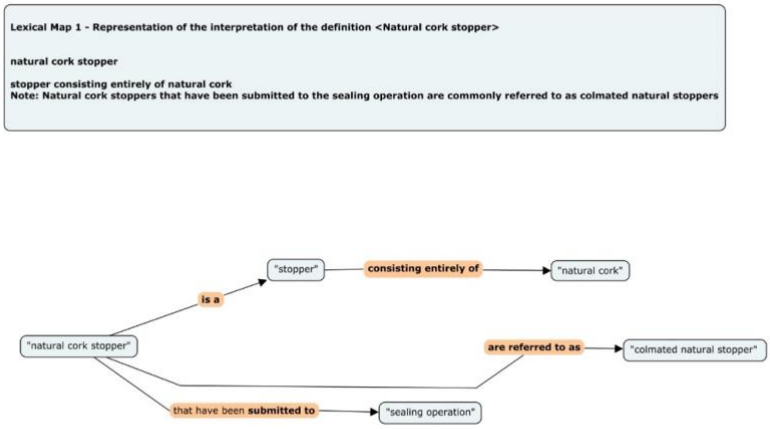

Figure 3. Lexical Map 1 – Representation of the text interpretation.

The linguistic analysis of 'Natural cork stopper' is visually presented in the form of lexical maps (Figure 3 and Figure 4). Lexical Map 1 (Figure 3) embodies the text interpretation, while Lexical Map 1.1 (Figure 4) delineates the lexical-semantic relations inferred from the lexical markers, coupled with the meanings of the related terms. In order to mirror the text structure of the definition for <Natural cork stopper>, Lexical Map 1 - Figure 3 is bifurcated into two sections: (1) the primary statement and (2) the secondary statement presented as a note, encapsulating our text interpretation. However, from our perspective, this particular definition poses challenges for knowledge organisation as two different objects are being defined. Typically, a single definition should define a single concept. Yet, each of the two statements within this definition describes a distinct variation of the <Natural cork stopper>. This observation will be revisited toward the conclusion of Section 5.

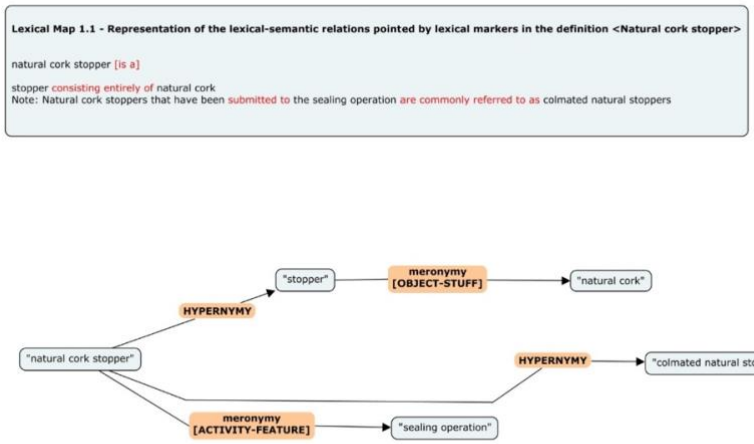

Figure 4. Lexical Map 1.1 – Representation of the lexical-semantic relations inferred from the lexical markers.

These lexical maps serve an important function by showcasing the terms actually related by a given lexical marker along the syntagmatic axis. This becomes especially significant when one of the paired terms is not directly within the co-text of the lexical marker but is found further along the text continuum, as exemplified by 'natural cork stopper' [TERM] 'are referred to as' [LEXICAL MARKER] 'colmated cork stopper' [TERM], in Lexical Map 1 - Figure 3.

5. The conceptual analysis

The conceptual analysis represents the subsequent phase in the analysis of the definition for <Natural cork stopper>. The *differential characteristics* [21] found within this definition are expressed by /natural cork/, /natural/, /colmated/, and /sealing operation/. The insights derived from this analysis have been organised and structured in Table 8, drawing upon the previously identified lexical markers from the linguistic analysis centred on this specific definition. Simultaneously, based on the linguistic interpretation of the data, we extrapolated conceptual relation identifiers.

CONCEPTUAL DIMENSION

Conceptual analysis of the definition for <Natural cork stopper> based on the Aristotelian formula X=Y+DC: X [species] = Y [genus] + DC [differential characteristic]

Analysis	Conceptual relation identifier	Conceptual relation	Interpretation	Transcription in X=Y+DC	Differential characteristic
natural cork stopper [is a] stopper	*is_a* [corresponds to LM *'is a'*]	natural cork stopper [is a] stopper	stopper [GENUS] natural cork stopper [SPECIES]	natural cork stopper [is a] stopper	
rolha de cortiça natura [é uma] rolha	*é_uma* [corresponds to LM 'é uma']	rolha de cortiça natura [é uma] rolha	rolha [GENUS] rolha de cortiça natural [SPECIES]	rolha de cortiça natura [é uma] rolha	
natural cork stopper [is made of] natural cork	*has_substance [corresponds to LM 'consisting entirely of']*	*natural cork stopper [is made of] natural cork*	*natural cork stopper [PRODUCT]* *natural cork [RAW MATERIAL]*	*natural cork stopper [is made of] natural cork*	*/natural cork/*
rolha de cortiça natural [é feita de] cortiça natural	*da_substância [corresponds to LM 'totalmente constituída de]*	*rolha de cortiça natural [é feita de] cortiça natural*	*rolha de cortiça natural [PRODUCT]* *cortiça natural [RAW MATERIAL]*	*rolha de cortiça natural [é feita de] cortiça natural*	*/cortiça natural/*
natural cork stopper [is made of] natural cork	*has_substance* [corresponds to LM 'consisting entirely of']	natural cork stopper [is made of] natural cork	cork [MATTER] natural [PROPERTY]	natural cork stopper [is made of] natural cork	*/natural/*
rolha de cortiça natural [é feita de] cortiça natural	*da_substância [corresponds to LM 'totalmente constituída de]*	*rolha de cortiça natural [é feita de] cortiça natural*	*cortiça [MATTER]* *natural [PROPERTY]*	*rolha de cortiça natural [é feita de] cortiça natural*	*/natural/*

natural cork stopper [is submitted to] sealing operation	*has_process* [corresponds to LM *'submitted to'*]	natural cork stopper [is submitted to] sealing operation	sealing operation = [PROCESS] ? = [RESULT]	natural cork stopper [is submitted to] sealing operation	/*sealing operation*/
rolha de cortiça natural [é submetida a] operação de colmatagem	*tem_processo* [corresponds to LM 'submetida a'	rolha de cortiça natural [é submetida a] operação de colmatagem	operação de colmatagem = [PROCESS] ? = [RESULT]	rolha de cortiça natural [é submetida a] operação de colmatagem	/operação de colmatagem/
colmated natural stopper [is a] natural cork stopper	*is_a* [corresponds to the LM '*commonly referred as*']	colmated natural stopper [is a] natural cork stopper	natural cork stopper [GENUS] colmated natural stopper [SPECIES]	colmated natural stopper [is a] natural cork stopper	/*colmated*/
rolha natural colmatada [é uma] rolha de cortiça de natural	*é_uma* [correspondes to LM 'comummente designada por']	rolha natural colmatada [é uma] rolha de cortiça de natural	rolha de cortiça natural [GENUS] rolha natural colmatada [SPECIES]	rolha natural colmatada [é uma] rolha de cortiça de natural	/colmatada/

Table 8. Conceptual analysis of the definition of <Natural cork stopper>.

As systematised in Table 8, we propose three conceptual relation identifiers: (1) *has_substance,* (2) *is_a*, and (3) *has_process*.

1. has_substance is conveyed through the lexical marker 'consisting entirely of' [*totalmente constituída por*], emphasising the substance constituting the object. As revealed in our linguistic analysis, the term 'natural cork' signifies the substance, denoting the material from which a particular object is crafted. Considering that a <Stopper> represents an object formed from a substance, we propose the conceptual relation identifier '*has_substance*' to represent this semantic relation. This relational aspect mirrors a pragmatic association—such as a thematic relation rooted in virtue or experience, or a dependency between concepts established through temporal or spatial proximity [22]. Here, a <Stopper> is a [PRODUCT] derived from a substance, specifically a [RAW MATERIAL]. From the interpretation of this information, we assume that an associative conceptual relation, subtype PRODUCT – RAW MATERIAL, is in place, where 'stopper' aligns with the meaning of PRODUCT, and 'natural cork' embodies the essence of RAW MATERIAL. This interpretation can be represented as follows:

(1) [stopper] PRODUCT *has_substance* [natural cork] RAW MATERIAL

The dichotomy of PRODUCT – RAW MATERIAL holds a dual significance in this phase of conceptual analysis: firstly, it forms the basis of the associative relation subtype, and secondly, it integrates into the Aristotelian formula[32][28] represented as X = Y + DC. In this formula, X denotes the specific concept, Y signifies the genus, and DC represents the differential characteristics. Using this formula aims to pinpoint, for the purpose of concept modeling, the characteristics outlined within the definition under analysis. To apply this formula effectively, it is imperative to identify two fundamental concepts: the specific concept and its genus.

2. ***is_a***

From our earlier linguistic analysis, we determined that 'natural cork stopper' is the specific term, while 'stopper' is the generic one. This inference leads us to conclude that the meaning of the concept <Natural Cork Stopper> is more specific than that of <Stopper>. Consequently, the established conceptual relation between these two concepts aligns with subsumption—a hierarchical relationship wherein a broader generic concept (*genus*) encompasses specific concepts (*species*). Thus, <Natural Cork Stopper> is the subordinate concept, labelled as [SPECIES], while <Stopper> is the superordinate concept, labelled as [GENUS]. This assumption can be represented as:

(2) [natural cork stopper] SPECIES is_a [stopper] GENUS

Having established the genus and species, we can now integrate these two elements into the formula X SPECIES = Y GENUS + DC, where:

(3) X = natural cork stopper; Y = stopper

In the subsequent stage, we identify the differential characteristic. Considering the information represented in (1), we can infer the following:

(4) X [natural cork stopper] = Y [stopper] + DC [natural cork]

The results obtained from the conceptual analysis serve as coordinates to systematise the knowledge encapsulated within the definition of <Natural cork stopper>, in the form of a conceptual map.

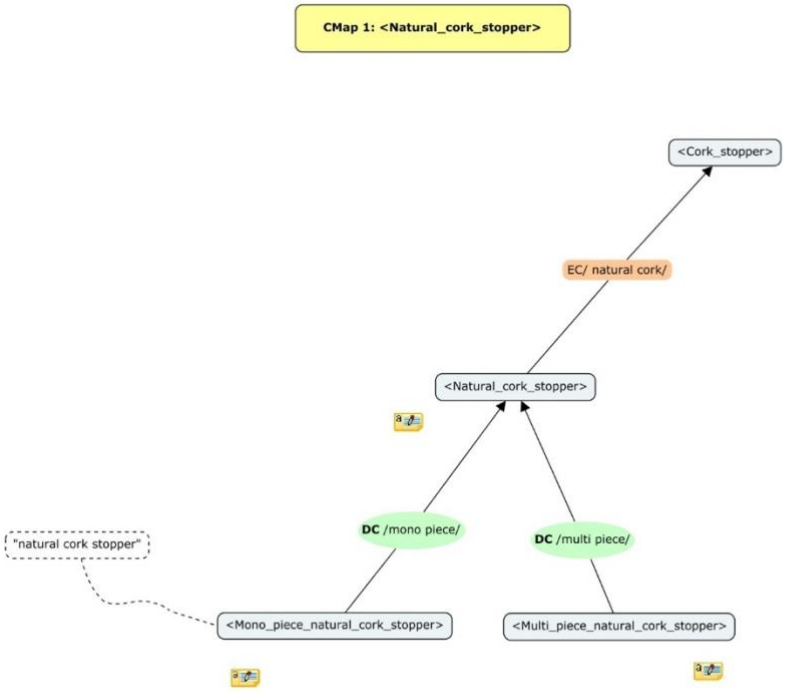

Figure 5. Conceptual Map 1 – Representation of <Natural_cork_stopper>, systematised by specific differentia, in CmapTools.

Conceptual Map 1 (Figure 5) is the conceptual representation of the first statement within the focused definition, from which we have inferred that a <Natural cork stopper> *is_a* <Cork stopper>. The concepts are systematised by *specific differentia*, based on the differential characteristics (DC) we have identified. As illustrated in Figure 5, the DC /natural cork/ underlies one of the *subdivision criteria* [21]. In this map, two axes of analysis are considered: Substance and Parts (the 'Parts' axis was inferred from Definition 1, Table 6). The conceptual information represented here, namely the Substance and Parts axes—founded upon the characteristics /natural cork/, /mono piece/ and /multi piece/—will function as key coordinates in the formulation of the formal description of the concept `NaturalCorkStopper` in Protégé. [12] The descriptors (identifiers) allocated to the concepts, such as <Multi_piece_natural_cork_stopper>, aim to reflect the characteristics of the concept. This rationale is further illustrated below:

[12] A Stanford University project that follows the recommendations of OWL 2 Web Ontology Language and Resource Description Framework (RDF) specifications from the World Wide Web Consortium (W3C). [https://protege.stanford.edu/][Accessed 2024-10-16].

(5) <Natural_cork_stopper> + /mono piece/ = <Mono_piece_natural_cork_stopper>

(6) <Natural_cork_stopper> + /multi piece/ = <Multi_piece_natural_cork_stopper>

3. ***has_process***

This last conceptual relation identifier stems from the information: <Natural cork stopper> 'is submitted to' /sealing operation/. The conceptual relation identifier *has_process* corresponds to the lexical marker 'submitted to'. Based on the meaning of the lexical marker—that clearly expresses an action—, the conceptual relation identifier intends to mirror the semantic dependency established between <Natural cork stopper> and the operation of sealing. The semantic dependency here observed falls under the associative relation of [PROCESS-RESULT] since 'sealing operation' points to a process. However, <Natural cork stopper> does not point to a result, but an object that undergoes a process:

(7) [?] RESULT *has_process* [sealing operation] PROCESS

The meaning of the RESULT emerged subsequently, drawn from the information conveyed by the characteristic /colmated/. The differential characteristic /colmated/ was deduced after the identification of the [GENUS-SPECIES] relation established between the concepts designated by 'colmated natural stopper' and 'natural cork stopper'. This conceptual relation was inferred from the lexical marker 'commonly referred as', indicating a specification between the concepts designated by those terms. This interpretation is represented as:

(8) [colmated natural stopper] SPECIES *is_a* [natural cork stopper] GENUS

and further transcribed into the formula X=Y+DC as follows:

(9) [colmated natural stopper] SPECIES = [natural cork stopper] GENUS + [colmated] DC

At this point, from the knowledge obtained in (9), we can finally fill in the information represented in (7), where the element indicating a RESULT was absent [?]:

(10) [colmated natural stopper] RESULT has_process [sealing operation] PROCESS

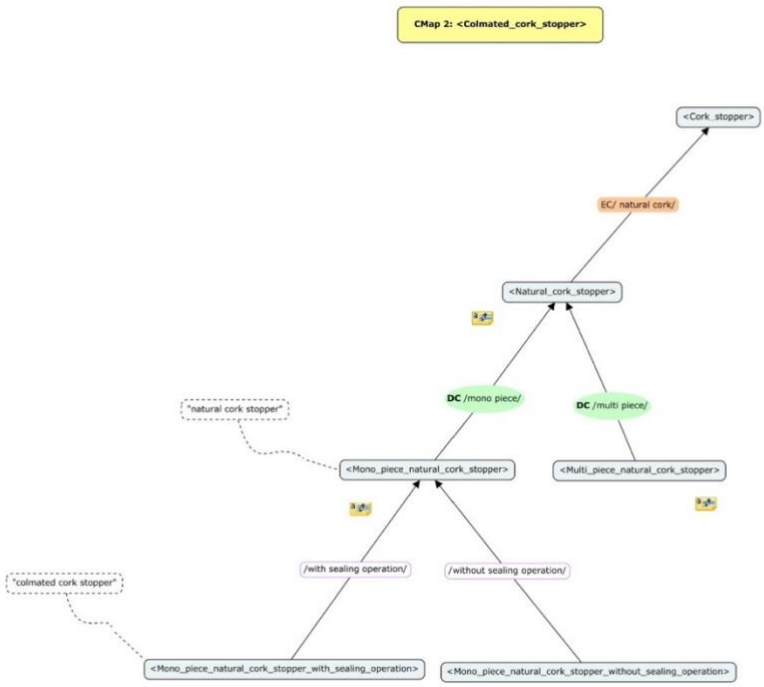

Figure 6. Conceptual Map 2: <Mono_piece_natural_cork_stopper_with_sealing_operation>.

Conceptual Map 2 (Figure 6) illustrates the conceptual analysis of the two sentences within the focused definition. Thus, three axes of analysis—Substance, Parts, and Finishing Processes—have been taken into account, extended by the characteristics /with sealing operation/ and /without sealing operation/. As depicted in Conceptual Map 2, these characteristics have led us to a distinct conceptual level. Specifically, the concept <Mono_piece_natural_cork_stopper_with_sealing_operation>, verbally designated as 'colmated cork stopper', emerges as a specialisation of the concept <Mono_piece_natural_cork_stopper>, designated as 'natural cork stopper'. Hence, these two concepts should not be considered on an equal basis, nor should they be defined within the same contextual framework, whether in natural language or formal languages.

5.1 Typology of definitional contexts

The conceptual relations inferred from the analysis of the lexical markers observed in the first five definitions (refer to Table 6) are summarised in Table 9.

Lexical marker	Conceptual relation identifier	Conceptual relation	Typology of definitional contexts governed by the DC
'is a' *(é uma)*	*is_a*	**SUBSUMPTION**	stopper [SPECIES]= product [GENUS] + [any DC added to the genus]
'*commonly referred as'*	*is_a*	**SUBSUMPTION**	colmated natural stopper [SPECIES] = natural cork stopper [GENUS] + colmated [DC added to the genus]

(commumente designadas por)			
'is a' *(é uma)*	*is_a*	**SUBSUMPTION**	colmated natural cork stopper [SPECIES] = stopper [GENUS] + [any DC added to the genus]
'intended to' *(destinado a)*	*has_function*	**ASSOCIATIVE**	stopper [SPECIES] = product [GENUS] + to seal bottles [FUNCTION=DC]
'obtained from' *(obtida de)*	*has_raw_material*	**ASSOCIATIVE**	stopper [SPECIES] = product [GENUS] + natural cork [SUBSTANCE=DC]
'obtained from' *(obtida de)*	*has_raw_material*	**ASSOCIATIVE**	stopper [SPECIES] = product [GENUS] + agglomerated cork [SUBSTANCE=DC]
'obtained from *(obtida de)*	*has_substance*	**ASSOCIATIVE**	natural cork [SPECIES] = cork [GENUS] + natural [SUBSTANCE=DC]
'obtained from' *(obtida de)*	*has_substance*	**ASSOCIATIVE**	natural cork [SPECIES] = cork [GENUS] + agglomerated [SUBSTANCE=DC]
'intended to' *(detinado a)*	*has_function*	**ASSOCIATIVE**	stopper [SPECIES] = piece of cork [GENUS] + to seal containers [FUNCTION=DC]
'piece of' *(peça de)*	*has_substance*	**ASSOCIATIVE**	stopper [SPECIES] = piece [GENUS] + cork [SUBSTANCE=DC]
'usually cylindrical' *(em geral cilíndrica)*	*has_shape*	**ASSOCIATIVE**	stopper [SPECIES] = piece of cork [GENUS] + cylindrical [SHAPE=DC]
'usually conical' *(em geral cónica)*	*has_shape*	**ASSOCIATIVE**	stopper [SPECIES] = piece of cork [GENUS] + conical [SHAPE=DC]
'usually prismatic' *(em geral prismática)*	*has_shape*	**ASSOCIATIVE**	stopper [SPECIES] = piece of cork [GENUS] + prismatic quadrangular [SHAPE=DC]
'sometimes with' *(por vezes de)*	*has_process*	**ASSOCIATIVE**	stopper [SPECIES] = piece of cork [GENUS] + rounded edges [PROCESS=DC]
'sometimes with' *(por vezes de)*	*has_process*	**ASSOCIATIVE**	stopper [SPECIES] = piece of cork [GENUS] + chamfered edges [PROCESS=DC]
''consisting entirely of' *(totalmente constituída por)*	*has_substance*	**ASSOCIATIVE**	natural cork stopper [SPECIES] = stopper [GENUS] + natural cork [SUBSTANCE=DC]
'consisting entirely of' *(totalmente constituída por)*	*has_substance*	**ASSOCIATIVE**	natural cork [GENUS] = cork [GENUS] + natural [SUBSTANCE=DC]
'submitted to' *(submetida a)*	*has_process*	**ASSOCIATIVE**	? [SPECIES] = natural cork stopper [GENUS] + sealing operation [DC]
'is made of' *(é feita de)*	*has_raw_material*	**ASSOCIATIVE**	colmated natural cork stopper [SPECIES] = stopper [GENUS] + natural cork [SUBSTANCE=DC]
'is made of' *(é feita de)*	*has_substance*	**ASSOCIATIVE**	colmated natural cork stopper [SPECIES] = natural cork stopper [GENUS] + colmated [SUBSTANCE=DC]
'its lenticels are filled' *(são obturadas as suas lenticelas)*	*has_process*	**ASSOCIATIVE**	colmated natural cork stopper [SPECIES] = natural cork stopper [GENUS] + filled lenticels [PROCESS=DC]

'results from' *(resulta de)*	*has_process*	**ASSOCIATIVE**	cork powder [SPECIES] = natural cork [GENUS] + dimensional finishing process [PROCESS=DC]
'consisting of' *(constituída por)*	*has_part*	**PARTITIVE**	stopper [SPECIES] = product [GENUS] + one piece [PARTS=DC]
'obtained from' *(obtida de)*	*has_part*	**PARTITIVE**	stopper [SPECIES] = product [GENUS] + several pieces [PARTS=DC]
'consisting of' *(contituída por)*	*has_part*	**PARTITIVE**	stopper [SPECIES] = piece of cork [GENUS] + one element [PARTS=DC]
'consisting of' *(constituída por)*	*has_part*	**PARTITIVE**	stopper [SPECIES] = piece of cork [GENUS] + several elements [PARTS=DC]

Table 9. Overview of the conceptual relations inferred from lexical markers.

As indicated in Table 9, the conceptual relation of subsumption is defined by the inclusion of any differential characteristic within a given definition, following the structure outlined in an intensional definition. This structure implies that by adjoining information within the intension of the [GENUS], the additional insights provided about the concept's position in the conceptual system are uniquely hierarchical. The associative relation operates similarly, albeit incorporating various other analytical axes. In this scenario, DC share semantic labels within a broader spectrum, specifically [SUBSTANCE], [FUNCTION], [PROCESS], and [SHAPE], due to the diverse and prolific semantic relations identified among concepts.

6. Building the ontology: OntoCork

OntoCork is an ontology—in the sense of [16]— in which the concepts of the cork domain are formally defined with logical constructs in OWL 2 Web Ontology Language Manchester Syntax[18][17]. Considering the extensive scope of the cork domain, we have focused on the set of concepts documented in Table 6 (Section 3.2). Protégé v.5.6.3. [29] served as the primary editor for constructing this ontology. The descriptive domain properties—or conceptual relations—we have formalised to underpin the ontology stem from the five retained axes of analysis: Function, Substance, Parts, Finishing Process, and Shape, as detailed in Table 10.

Axis of analysis	**Format in Protégé**	**Type of conceptual relation**
FUNCTION	hasFunction	Associative relation, subtype [OBJECT-FUNCTION]
SUBSTANCE	IsMadeOf	Associative relation, covering both subtypes [RAW MATERIAL – PRODUCT] and [MATTER/SUBSTANCE – PROPERTY]
PARTS	hasStructure	Partitive relation [PART-WHOLE]
FINISHING PROCESS	hasFinishingProcesses	Associative relation, within the subtype [PROCESS-RESULT]
SHAPE	hasShape	Associative relation, subtype [OBJECT-SHAPE]

Table 10. Five core conceptual relations of the ontology.

In Protégé, conceptual relations are expressed as `owl:ObjectProperties`, adhering to OWL[13] terminology, which we formalised as `hasStructure`, `hasFunction`, `hasShape`, `hasFinishingProcess`, and `isMadeOf` for the ontology construction. As outlined in Table 10, these conceptual relations correspond to the partitive and associative relations, playing a pivotal role in the process of formally defining concepts since the enumeration of the characteristics that build up a given concept is through them expressed.

Our ontology construction started with representing the five axes of analysis as generic concepts, whose individuals serve to assert object values— e.g. `Cylindrical`—through the types (and subtypes) of the core conceptual relations (see Table 10).

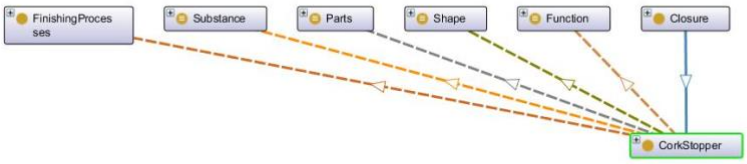

Figure 7. Representation, in OntoGraf, of the conceptual relations used to define <CorkStopper>.

Figure 7 is a representation of the conceptual relations used to define <CorkStopper>, in OntoGraf.[14] As represented by the coloured dashed arcs, the concept `CorkStopper` relates with `Parts`, `Substance`, `Shape`, `FinishingProcesses`, and `Function`. The different colours of the arcs represent different types of relations, e.g., the orange dashed arc between `CorkStopper` and `Substance` represents the owl:ObjectProperty `IsMadeOf`, which corresponds to the associative relation subtype [MATTER/SUBSTANCE-PROPERTY] in our study. Finally, the subsumption relation is denoted by the solid blue arc, illustrating that `CorkStopper` is a subtype of `Closure`.

6.1 The case of <WashedMonopieceNaturalCorkStopper>, a type of <Semi-finishedStopper>

We will focus on `WashedMonopieceNaturalCorkStopper` to demonstrate our methodology for describing the concepts' characteristics in Protégé, adhering to the structure of an intensional definition.

[13] https://www.w3.org/TR/owl-ref/ [Accessed 2024-10-16]

[14] https://protegewiki.stanford.edu/wiki/OntoGraf [Accessed 2024-10-16].

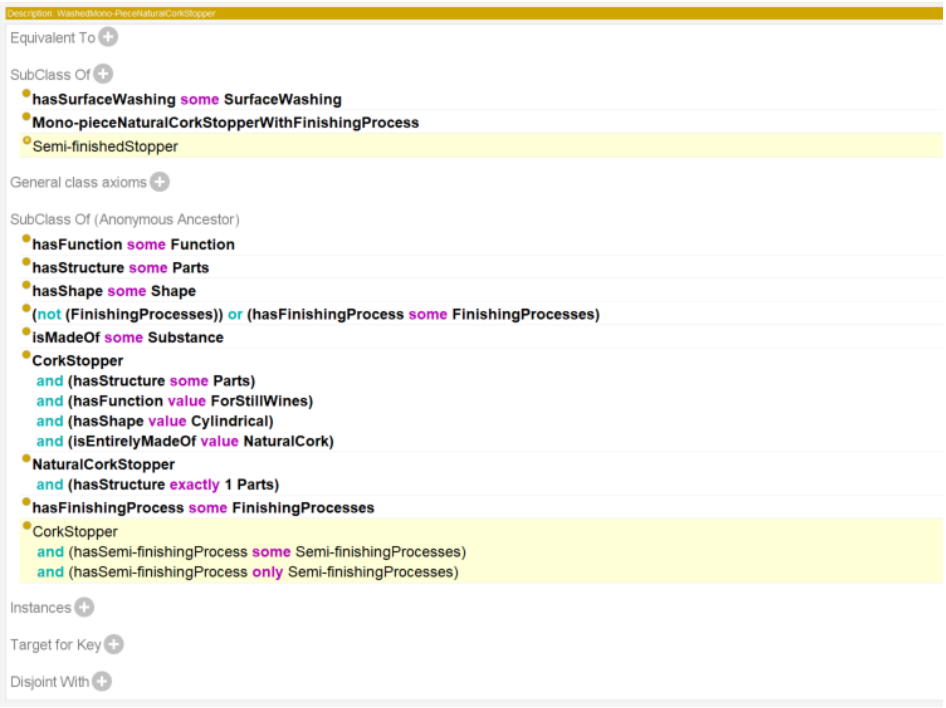

Figure 8. The characteristics of <WashedMonopieceNaturalCorkStopper>, in Protégé.

Figure 8 represents the concept editor window within Protégé, where the description of the concept is implemented in OWL 2 Manchester Syntax. As previously noted, these characteristics are expressed by `owl:ObjectProperties`, aligning with the conceptual relations in this study. The intensional definition of `WashedMonopieceNaturalCorkStopper` is asserted under 'SubClassOf' as follows:

(1) `SubClassOf:`
`hasSurfaceWashing some SufaceWashing`
`Mono-pieceNaturalCorkStopperWithFinishingProcess`

which can be (informally) represented as:

(2) <WashedMonopieceNaturalCorkStopper> *is_a*
<MonopieceNaturalCorkStopper>
<WashedMonopieceNaturalCorkStopper> *has_process* < SurfaceWahing>

Expression (2) clearly mirrors the structure of an intensional definition: $X_{[SPECIES]}=Y_{[GENUS]}+DC$. The extension of the concept comprises the remaining set of characteristics expressed under 'SubClass Of (Anonymous ancestor)' (see Figure 8), for they are inherited from the subsuming concepts we have previously defined, namely `NaturalCorkStopper`, `CorkStopper` and `Closure`. Lastly, highlighted in yellow (Figure 8), the reasoner HermiT[15] infers the following classification:

[15] HermiT – a plugin of Protégé (http://www.hermit-reasoner.com/) [Accessed 2024-10-16]

(3) <WashedMonopieceNaturalCorkStopper> *is a* <Semi-finishedStopper>

This classification (3) results from the constraints we have established to the domain and range of `hasSurfaceWashing` — a subtype of `hasSemi-finishingProcess` —, an owl:ObjectProperty used for expressing the characteristic /with washing treatment/. The constraints are represented in the following XML schema:

```
< ! -- Domain -->

<ObjectPropertyDomain>
        <ObjectProperty IRI="#hasSurfaceWashing"/>
        <Class IRI="#Semi-finishedStopper"/>
</ObjectPropertyDomain>

< ! -- Range -->

<ObjectPropertyRange>
        <ObjectProperty IRI="#hasSurfaceWashing"/>
        <Class IRI="#SurfaceWashing"/>
    </ObjectPropertyRange>
```

We have employed the same methodology across all subtypes of `has-SemifinishingProcess`, including the latter. Consequently, whenever a concept is defined with the characteristic denoted by each of these owl:ObjectProperty subtypes, the reasoner will classify it as a member of the class `Semi-finishedStopper`. This classification conveys a piece of information regarding the object's stage of completion within the manufacturing process.

Table 11showcases the ontology metrics and axioms that form the structure of OntoCork.

Metrics	
Axiom	791
Logical axiom	274
Declaration axioms	152
Classes	54
Object Property	40
Data property	9
Individual	24
Annotation property	29
Class axioms	
SubClassOf	52
EquivalentClasses	15
DisjointClasses	17
Object property axioms	
SubObjectPropertyOf	36
TransitiveObjectProperty	8
ObjectPropertyDomain	36
ObjectPropertyRange	34
Data property axioms	
SubDataPropertyOf	3

DataPropertyDomain	15
DataPropertyRange	4
Individual axioms	
ObjectPropertyAssertion	15
DatapropertyAssertion	9
DiferentIndividuals	1
Annotation axioms	
AnnotationAssertion	363

Table 11. The ontology metrics and axioms.

As documented in Table 11, the ontology comprises 54 classes (concepts) that align with various typologies such as types of <CorkStopper>, types of <FinishingProcess>, and the five axes of analysis, e.g. <Substance>. The Class axioms and Object property axioms, implemented to establish relationships among concepts and thereby define them, are illustrated in Figure 9.

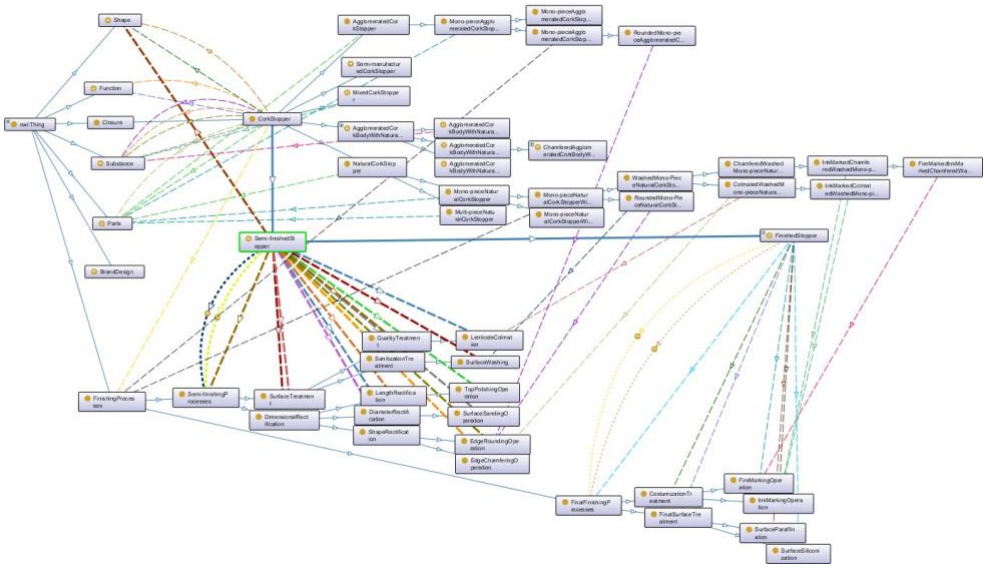

Figure 9. Representation of the conceptual relations, and corresponding restrictions, used to define types of <CorkStoppers> and their correlation with <FinishingProcesses>, in OntoGraf.

Figure 9 illustrates, in OntoGraf, a representation encompassing the concepts identified from the corpus definitions. The concepts are systematised either hierarchically, denoted by blue solid lines, or associatively (i.e., in a pragmatic dependency), represented by dashed lines, following the axiomatic constructs we have implemented to express their characteristics. For visualisation purposes, we have emphasised the concept `Semi-finishedStopper` to display its relationships with other concepts, such as the `WashedMonopieceNaturalCorkStopper`. As depicted by the blue arc, this concept is a subtype of `MonoPieceNaturalCorkStopperWithFinishingProcess`. The concepts `WashedMonopieceNaturalCorkStopper` and `SurfaceWashing` are related by `hasSurfaceWashing` reflecting the associative relation, subtype [PROCESS-RESULT]. This conceptual relation is established through the differential characteristic /with washing treatment/, drawn from the analysis of the concept `FinishingProcesses`. Thus,

`hasSurfaceWashing` is the associative relation that induces the specification of `MonopieceNaturalCorkStopperWithFinishingProcess` by specific differentia.

6.2 The ontology evaluation

The topic of evaluation is widely discussed within the knowledge engineering community. As highlighted in ontology literature, there exists a multitude of methods to evaluate an ontology[11][29][35][23]. However, given the early stage of OntoCork's development, we will consider the modelling of knowledge in Protégé with an embedded description logic reasoner as a preliminary evaluation form. This assertion stems from the accurate classifications made by the reasoner, specifically regarding the completion status of a given object, unveiling the coherence and consistency of the axiomatic structures we have implemented to construct the ontology. As depicted in Figure 10, these classifications manifest at either the individual or instance level.

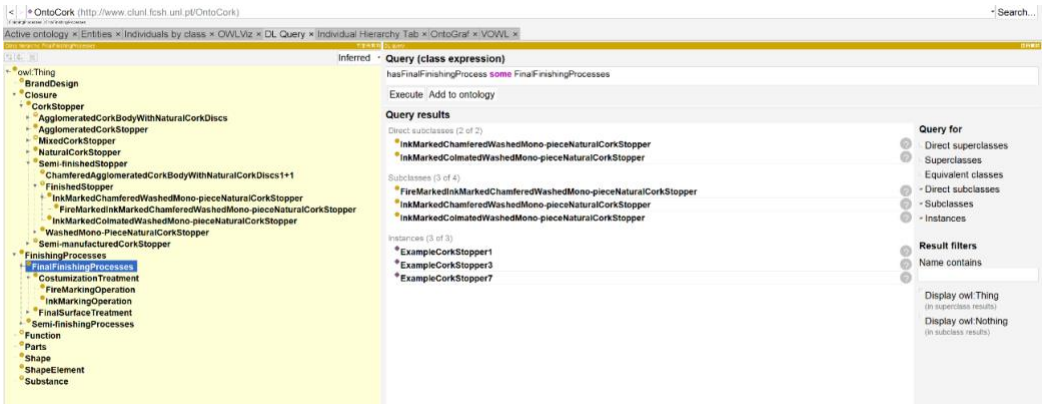

Figure 10. Results, in Protégé, to the competency question: Which stoppers have 'final finishing processes'?

Certain methodologies for ontology construction recommend devising use-case scenarios and competency questions [44][40]. To evaluate the ontology using use-case scenarios, one initially needs to define the ontology's purpose. In our context, we have observed a limited presence of cork-related domains in existing ontologies, except for the concepts <Wooden cork> — a type of <Stopper> — in the FoodOn Food[16] ontology and <Cork cambium> in the Plant Ontology[17]. The purpose behind constructing an ontology to express the knowledge from the cork transformation sector lies in the perspective of (i) introducing an innovative dimension in the domain, and (ii) building a tool tailored for experts, future experts and language specialists, where the shared understanding of this specific domain may be used as a unifying framework to solve, for instance, communication issues [44].

[16] https://purl.bioontology.org/ontology/FOODON?conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FFOODON_03490256 [Accessed 2024-10-16].

[17] https://purl.bioontology.org/ontology/PO?conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FPO_0005599 [2024-10-16].

OntoCork is a micro-domain ontology designed to address two main typologies: (1) the type of cork stoppers concerning the type of cork (raw material) from which they are manufactured; and (2) the type of operations pertaining to finishing processes. Ultimately, this ontology aims to classify the completion status of the cork stopper as a (semi)finished product, contingent upon the last operation it underwent. We employed DL Query—a Protégé plugin that enables users to elaborate queries using class expression constructs—to formulate competency questions aimed at evaluating OntoCork, such as the following examples:

a) Which stoppers have 'final finishing processes'?

b) What is a 'semi-finished stopper'?

The outcomes illustrated in Figure 10 for question (a) consistently identify eligible individuals and instances. Concerning question (b), while the results are accurate (not here illustrated for the economy of space), they may introduce some additional information: the presence of the generic <FinishedStopper> is included as a direct subclass, a deliberate decision made during the ontology's implementation. This choice allows for a seamless status transition, e.g., an instance initially classified as <Semi-finishedStopper> promptly shifts to a completed state once its description includes a restriction whose domain is constrained to <FinishedStopper>. Our assurance lies in the assumption that reasoner classifications are inferred from logical axioms. This validates the coherence of the constraints we have implemented, evident in the accurate classifications of the concepts depicted in Figure 11, visualised with the OWLViz[18] plugin.

[18] https://protegewiki.stanford.edu/wiki/OWLViz [Accessed 2024-10-16].

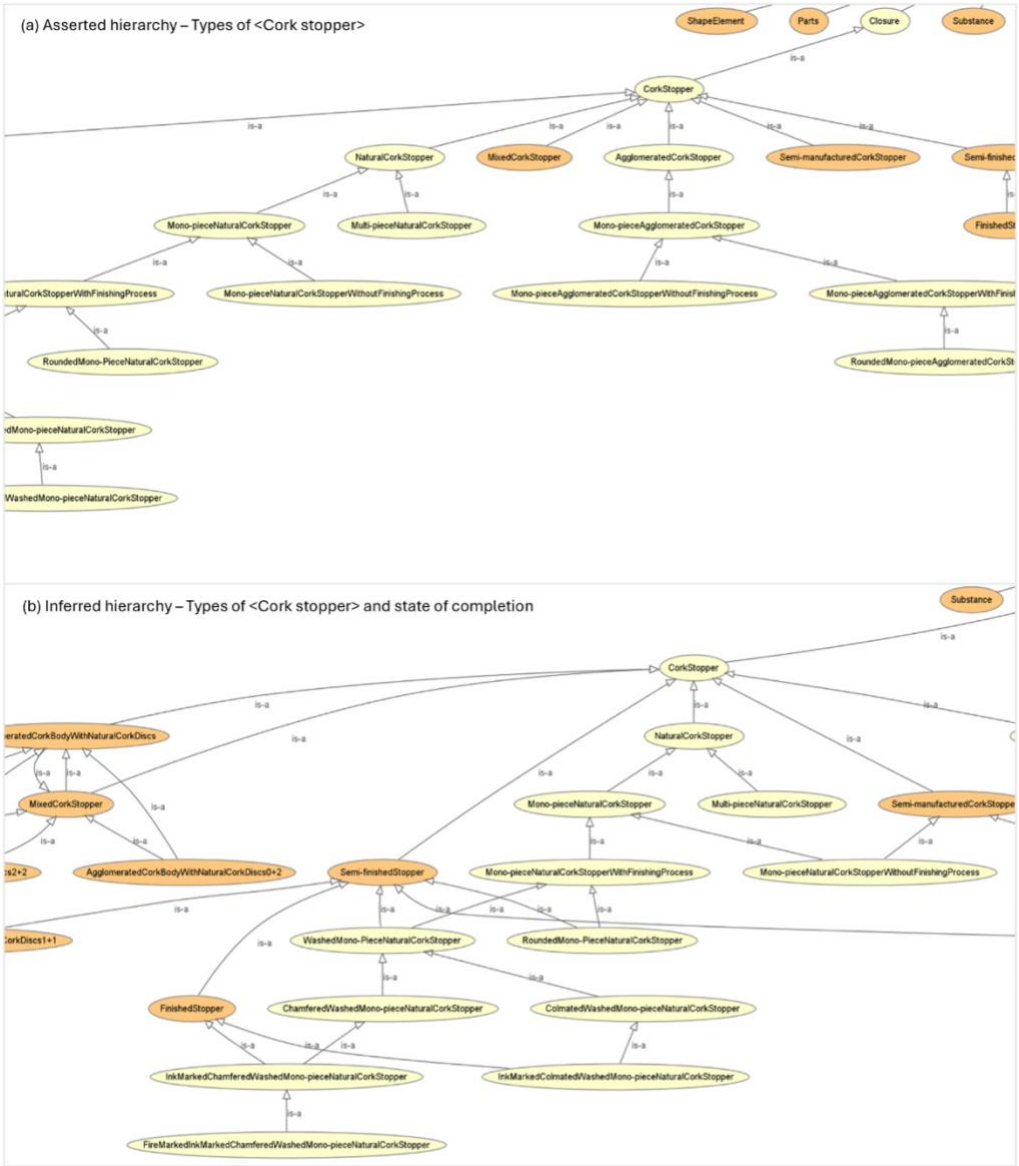

Figure 11. (a) Representation of the asserted definitions; (b) The reasoner classifications inferred from the logical axioms.

Figure 11 includes two frames, (a) and (b), each illustrating the hierarchical systematisation of concepts at different levels of conceptual granularity. This variation results from the formal definitions we have implemented in OWL 2 Manchester Syntax. Frame (a) shows an asserted hierarchy, representing concepts defined using genus-differentia. Frame (b) depicts an inferred hierarchy, where the same concepts are classified by the reasoner according to additional information, such as their completion stage, provided by the logical constructs (see Annex 1 for the whole hierarchy).

Hence, given that the outcomes align with our expectations, we conclude that the instances within this ontology effectively reflect coherent use-case scenarios, thus justifying the existence and properties of the objects within the ontology [44].

7. CorkTerm: the linguistic resource

The primary aim of this study is to create a digital dictionary designed as a multilingual and multimodal product, where several resources, namely linguistic, conceptual, and multimedia, are paired to facilitate user knowledge acquisition. To construct the dictionary, we used Lexonomy[19], a publishing dictionary editor.

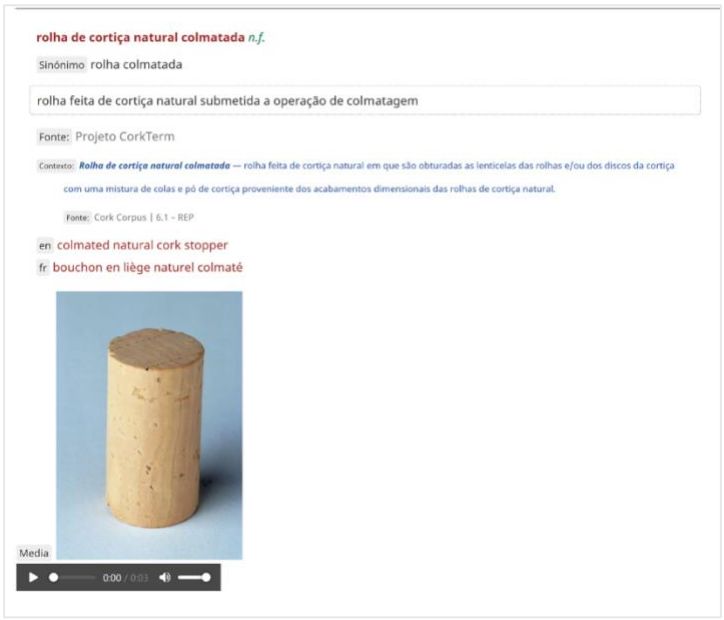

Figure 12. The term "*rolha de cortiça natural colmatada"* (colmated natural cork stopper) in the dictionary editor Lexonomy.

Figure 12 illustrates the terminological entry for the term *'rolha de cortiça natural colmatada'* (colmated natural cork stopper), classified as a feminine noun denoted by the descriptor 'n. f.'. This term is considered the preferred term. Directly below, an alternative form is presented, namely '*rolha colmatada'* (colmated stopper), labelled as '*Sinónimo'* (synonym). This second descriptor is aligned with the skos:altLabel used in Protégé. The decision to include this alternative form stemmed from its frequency in the corpus and the knowledge it references. These observations led us to infer that the alternative form holds reduced information compared to the preferred term. Thus, if searched for within the linguistic resource, the definition of the concept can be obtained through this alternative form.

[19] https://www.lexonomy.eu/ [Accessed on 2024-10-16].

Within the microstructure of the terminological record for '*rolha de cortiça* natural *colmatada*' (colmated natural cork stopper), examples showcasing the term's usage in texts are provided, alongside their respective sources. The equivalents in English and French are identical to those we have assigned to the ontology's concepts, as rdfs:label annotations.

7.1 Linking the terminological resources

The ongoing terminological dictionary, currently being developed in Lexonomy, is accessible from the ontology through the AnnotationAssertion property in Protégé, in this case, expressed within the skos:definition[20]. As an initial step, we used the FOAF Vocabulary Specification 0.99[21], sourced from the Linked Open Vocabularies (LOV) website[22], to embed the dictionary's URI. FOAF also serves as the vocabulary for linking images to concepts in OntoCork. Lastly, Lexonomy's interface enables users to link dictionary entries to external resources, in our case CorkTerm and OntoCork (a work in progress). This interoperability among digital resources represents a significant advantage for democratising knowledge, particularly aligning with the FAIR principles [46].

8. Discussion

8.1 Issues raised by the definitions

Using definitions written by experts as a basis has a positive impact on the results, for they ensure the quality of the domain ontology. As such, the quality of the ontology will facilitate the reduction of ambiguity in expert communication. The definitions we worked on are generally of homogeneous quality since they are retrieved from official documents. Still, there are definitions with less quality than others. To overcome the difficulties raised by this imbalance, we analysed several definitions for the same concept, including definitional contexts, so that we could gather the information that enables us to infer the underlying conceptual organisation.

8.2 Text-mining methods in the terminological work

Using text-mining methods to collect data in order to build ontologies is at the base of our research in terminology. Our methods align with the concept of Information Extraction (IE), particularly in the keyword search phase [39]. It is essential to emphasize that the outcomes derived from our methods do not aim to establish an automatic IE system for extracting information from texts. Instead, our methodology aims to add a qualitative aspect to the interpretation of automatically obtained results. For instance, when identifying linguistic expressions denoting hypernymy, we address concerns highlighted in studies like [26]. These studies suggest that approaches reliant on statistical measurements in artificial intelligence might capture noise and idiosyncratic data. Consequently, the challenge arises in converting implicit knowledge from text into a formalised ontology representation, facing difficulties concerning both the quantity and quality of information.

[20] https://www.w3.org/2009/08/skos-reference/skos.html#definition [Accessed 2024-10-16].

[21] http://xmlns.com/foaf/0.1/ [Accessed 2024-10-16].

[22] https://lov.linkeddata.es/dataset/lov/ [Accessed 2024-10-16].

Given the intricate nature of natural language, identifying associative relations—perceived as inherently conceptual from our standpoint [21]—proved equally complex in this study. In the domain of cork, where concepts pertaining to activities form the core of the cork transformation sector, diverse linguistic patterns convey associative relations. To address this challenge, we developed an iterative approach while exploring corpora with CQL queries. This method enabled us to extract linguistic data indicative of relationships beyond hypernymy. The results obtained from the methodology presented in this study prompt us to hypothesise that our methods might be applicable to other languages within the same domain. This assumption lays the groundwork for future research.

8.3 Future work

In our forthcoming endeavours, we aim to map the conceptual and linguistic information contained in our developed resources—the ontology, corpus, and the ongoing terminological dictionary in Lexonomy—as linked data using interoperable Linked Open Vocabularies. Considering that dictionary entries in Lexonomy are stored as XML documents [27], we envision its potential alignment with TEI-Lex0[23], a technical specification and a set of community-based recommendations tailored for encoding machine-readable dictionaries. Additionally, we seek alignment with the vocabulary of the Lexicon Model for Ontologies (*OntoLex-lemon*) [31], which provides a 'vocabulary that allows ontologies to be enriched with information about how the vocabulary elements described in them are realized linguistically, in particular in natural languages' [31]. These functionalities and inherent interoperability underlie our motivation to further develop our project. We find the lexicon core model OntoLex-lemon well-suited for modelling the linguistic information encapsulated in the ontology. This includes terms, terminological definitions, and equivalents annotated with RDF vocabularies, facilitating their searchability and reusability within the Semantic Web. Furthermore, the conceptual information derived from OntoCork's ontology elements (including classes, properties, individuals, and so forth) is expected to achieve increased expressivity. Based on the premise that '[r]ich linguistic grounding [for ontologies] includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or vocabulary' [31], we speculate its potential usefulness in illustrating the classification of the object <CorkStopper> as a finalised or non-finalised product. This feature could thereby provide a second sense within the dictionary entry.

At the time of submitting this article, the Ontology Lexicon community group was developing a Terminology module extension for the *Lexicon Model for Ontologies*. Therefore, this article does not include discussions of this module due to its evolving status during this period.

Lastly, we aim to extend OntoCork and align it with larger ontologies, such as FoodOn Food ontology, available in BioPortal.[24]

[23] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html [Acessed 2024-10-16].

[24] https://bioportal.bioontology.org/ [Accessed 2024-10-16].

Conclusion

Through this research, our aim was to elucidate the method employed in constructing an ontology derived from human analyses of linguistic data. We sought to analyse linguistic and conceptual aspects as distinct yet interconnected phenomena. Texts serve as conduits for knowledge transfer, enabling us to extract concept characteristics linguistically expressed through lexical markers indicating lexical-semantic relations. This approach effectively captured domain-specific conceptual relations via the formula X=Y+DC.

As demonstrated in this study, we were able to propose a preliminary conceptual organisation of the subject field. We have bridged three main aspects in our study: (i) classical aspects of Aristotelian logic; (ii) the methodology underlying our terminological work, where characteristics play a pivotal role in the analysis or the drafting of intensional definitions; and (iii) formal definitions. To achieve this, we employed Protégé and the inherent OWL 2 Web Ontology Language to formally describe domain concepts, establishing relationships via abstract syntaxes and enabling formal reasoning. This process ensures consistent definitions through the soundness of the 'reason-able' ontology.

Acknowledgments

The research is supported by the Portuguese national funding through the FCT – Portuguese Foundation for Science and Technology, I.P. as part of the project UIDB/LIN/03213/2020; 10.54499/UIDB/03213/2020 and UIDP/LIN/03213/2020; 10.54499/UIDP/03213/2020 – Linguistics Research Centre of NOVA University Lisbon (CLUNL).

This paper is partially based on a Ph.D. thesis accomplished within the scope of a co-tutelle agreement between the Universidade NOVA de Lisboa and the Université Savoie Mont Blanc http://hdl.handle.net/10362/111722 ; [HAL Id: tel-03106436, version 1]. Although using the same data and methodology presented in the conference article "Extracting knowledge-rich information from definitions. A corpus-based approach to build a conceptual-based terminological resource." (MDTT 2023. Lisboa: CEUR Workshop Proceedings (CEUR-WS.org)), the present paper displays different outcomes and the methods are thoroughly presented.

This work was conducted using Protégé.

References

[1] Agbago, Akakpo, and Caroline Barrière. "Corpus Construction for Terminology." *Corpus Linguistics 2005 Conference*. Birmingham: National Research Council of Canada, 2005.

[2] APCOR. *Cortiça, Cork2013*. Brochura; Pdf. Santa Maria de Lamas: APCOR, 2013.

[3] Atkins, Sue, Jeremy Clear, and Nicholas Ostler. "Corpus Design Criteria." *Literary and Linguistic Computing* 7, no. 1 (1992): 1 - 16. https://doi.org/10.1093/llc/7.1.1

[4] Baker, Paul, Andrew Hardie, and Tony McEnery. *A Glossary of Corpus Linguistics*. In the series Glossaries in Linguistics. Edinburgh University Press, 2006. https://doi.org/10.1515/9780748626908

[5] Barata, and Ganhão. *Caracterização Processual e Económico-Financeira do Subsector Transformador e Comercial das Rolhas de Cortiça Naturais e Aglomeradas*. Pdf. AIEC; FERA. Lisboa, 2004.

[6] Bicho, Margarida. *A Rolha de Cortiça: da floresta à utilização*. Santa Maria de Lamas: APCOR-Associação Portuguesa da Cortiça, 2004.

[7] Bowker, Lynne, and Jennifer Pearson. *Working with specialized language: a practicle guide to using corpora*. London: Routledge, 2002.

[8] Brezina, Vaclav. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press, 2018. https://doi.org/10.1017/9781316410899

[9] CIPR V5. *Código Internacional das Práticas Rolheiras*. 5ª versão. Pdf. APCOR / AIEC. Santa Maria da Feira / Lisboa: C.E.Liège, 2006.

[10] Costa, Rute. "Pressupostos teóricos e metodológicos para a extracção automática de unidades terminológicas multilexémicas." *PhD Thesis*. Lisboa: Universidade Nova de Lisboa, Faculdade de Ciências Sociais e Humanas, 2001.

[11] Fernández, Mariano, Asunción Gómez-Pérez, and Natalia Juristo. "Methontology: From Ontological Art Towards Ontological Engineering." *Ontological Engineering | AAAI Spring Symposium*. Association for the Advancement of Artificial Intelligence, 1997.

[12] Gil, Luís. *A Rolha de Cortiça e a sua Relação com o Vinho*. Portalegre: Agrupamento de Produtores Agrícolas e Florestais do Norte do Alentejo - APAFNA, 2002.

[13] Gil, Luís. *Cork as a Building Material | Technical manual*. Pdf. Santa Maria de Lamas: APCOR - Portuguese Cork Association, 2007.

[14] —. *Cortiça: produção, tecnologia e aplicação*. Lisboa: Instituto Nacional de Engenharia e Tecnologia Industrial, 1998.

[15] Gil, Luis. "Novas aplicações da cortiça." INGENIUM, 2015. 36-38.

[16] Gruber, Tom. "Ontology." In *Encyclopedia od Database Systems*, edited by Ling Liu and Tamer Özsu, 1963-1965. Boston, MA: Springer, 2009. https://doi.org/10.1007/978-0-387-39940-9_1318

[17] Horridge, Matthew, and Peter F. Patel-Schneider. "OWL 2 Web Ontology Language Manchester Syntax (Second Edition)." *W3C Working Group Note*. 11 December 2012. https://www.w3.org/TR/owl2-manchester-syntax/#ref-manchester-owl-dl.

[18] Horridge, Matthew, Nick Drummond, Goodwin, Rector, Alan John, Robert Stevens, and Hai H. Wang. "The Manchester OWL Syntax." *Proceedings of the OWLED*06 Workshop on OWL: Experiences and Directions*. Athens, Georgia, USA: dblp | computer science bibliography, 2006.

[19] INETI. *Guia Técnico Sectorial - Indústria da Cortiça.* Lisboa: Instituto Nacional de Engenharia e Tecnologia Industrial, 2001.

[20] INPI. "A utilização e a valorização da propriedade industrial no sector da cortiça." *Colecção leituras e propriedade industrial.* Vol. III. Pdf. Prod. Instituto Nacional da Propriedade Industrial. Lisboa: INPI, 12 2005.

[21] ISO/FDIS 1087 (E). "Terminology work and terminology science - Vocabulary." Suisse: ISO, 2019.

[22] ISO/NF 704. "Travail terminologique - Principes et méthodes." La Plaine Saint-Denis: Association Française de Normalisation (AFNOR), 2009.

[23] Izquierdo, Alba Fernández. *Themis.* 07 2020. http://themis.linkeddata.es/index.html.

[24] Laviosa, Sara. "Corpus Linguistics and translation studies." In *Perspectives on Corpus Linguistics*, edited by Vander Viana, Sonia Zyngier and Geoff Barnbrook, 131-153. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2011. https://doi.org/10.1075/scl.48

[25] L'Homme, Marie Claude. *La Terminologie: principes et techniques.* Collection : Paramètres. Montréal: Les presses de l'Université de Montréal, 2004. https://doi.org/10.4000/books.pum.10693

[26] Lim, Edward, James Liu, and Raymond Lee. *Knowledge Seeker – Ontology Modelling for Information Search and Management.* Series: Intelligent Systems Reference Library. Edited by Janusz Kacprzyk, Jain and Lakhmi. Hong Kong: Springer Berlin, Heidelberg, 2011.

[27] Mechura, Michal. "Introducing Lexonomy: an open-source dictionary writing." *Proceedings of the eLex 2017 conference*, 2017.

[28] Meyer, Ingrid. "Extracting Knowledge-Rich contexts for terminography: a conceptual and methodological framework." In *Recent Advances in Computational Terminology*, edited by Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme, 279 - 302. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2001. https://doi.org/10.1075/nlp.2.15mey

[29] Musen, M. A., and Protégé-team. "The Protégé project: A look back and a look forward." *AI Matters* 1, no. 4 (June 2015). https://doi.org/10.1145/2757001.2757003

[30] Norma Mínima V.1. "Guia Internacional de Compra de Rolhas para Vinhos Tranquilos." *A trabalhar com o Comércio (trade) e com a Indústria da Cortiça.* Versão 1. Pdf. Grupo de Utilizadores de Rolha de Cortiça NATURAL. 2007.

[31] Ontology-Lexicon Community Group. "Lexicon Model for Ontologies." *W3C Community Group Final Report.* Edited by Philipp Cimiano, John P. McCrae and Paul Buitelaar. 10 May 2016. https://www.w3.org/2016/05/ontolex/.

[32] Pearson, Jennifer. *Terms in context.* Amsterdam: John Benjamins Publishing Company, 1998. https://doi.org/10.1075/scl.1

[33] Pereira, Helena. *Cork: Biology, Production and Uses.* Amsterdam: Elsevier, 2007.

[34] Pottier, Bernard. *Théorie et analyse en Linguitique.* 2, corrigée. Paris: HACHETTE, Supérieur, 1992.

[35] Poveda-Villalón, María, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. "OOPS!(Ontology Pitfall Scanner!): An on-line tool for ontology evaluation." *International Journal on Semantic Web and Information Systems (IJSWIS)* (IGI Global) 10, no. 2 (2014): 7-34.

[36] Ramos, Margarida. "Knowledge Organization and Terminology: application to Cork." *PhD Thesis.* Lisboa: Universidade NOVA de Lisboa, Faculdade de Ciências Socias e Humanas; Université Savoie Mont Blanc, Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance, 2020. http://hdl.handle.net/10362/111722; https://hal.science/tel-03106436

[37] Ramos, Margarida. *OntoCork.* Dataset - OWL File. 2020. https://doi.org/10.34619/a27q-1ryd

[38] Ramos, Margarida, and Rute Costa. "Extracting knowledge rich information from definitions. A corpus-based approach to build a conceptual based terminological resource." *2nd International Conference on Multilingual Digital Terminology Today (MDTT 2023).* CEUR Workshop Proceedings (CEUR-WS.org), 2023.

[39] Ramzan, Talib, K. Hanif Muhammad, Ayesha Shaeela, and Fatima Fakeeha. "Text Mining: Techniques, Applications and Issues." *International Journal of Advanced Computer Science and Applications (IJACSA)* 7, no. 11 (2016): 414 - 418. https://dx.doi.org/10.14569/IJACSA.2016.071153

[40] Sabou, Marta, and Miriam Fernandez. "Ontology (Network) Evaluation." In *Ontology Engineering in a Networked World*, edited by M. Suárez-Figueroa, A. Gómez-Pérez, E. Motta and A Gangemi, 193-212. Berlin, Heidelberg: Springer, 2012.

[41] Suárez-Figueroa, Mari Carmen, Asunción Gómez-Pérez, and Mariano Fernández-López. "The NeOn Methodology for Ontology." In *Ontology Engineering in a Networked World*, edited by Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, Enrico Motta and Aldo Gangemi, 9-34. Berlin, Heidelberg: Springer, 2007. https://doi.org/10.1007/978-3-642-24794-1_9

[42] Taber, George. *To cork or not to cork.* New York: SCRIBNER, 2009.

[43] Tognini Bonelli, Elena. "Theoretical overview of the evolution of corpus linguistics." Chap. 2 in *The Routledge Handbook of Corpus Linguistics*, edited by Anne O'Keeffe and Michael McCarthy, 14-27. London: Routledge, 2010. https://doi.org/10.4324/9780203856949

[44] Uschold, Mike, and Michael Gruninger. "Ontologies: principles, methods and applications." *The Knowledge Engineering Review* 11 (1996): 93-136.

[45] Viana, Vander. "The politics of Corpus Linguistics." In *Perspectives on Corpus Linguistics*, edited by Vander Viana, Sonia Zyngier and Geoff Barnbrook, 229-245. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2011. https://doi.org/10.1075/scl.48

[46] Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data*, nº 3 (03 2016). https://doi.org/10.1038/sdata.2016.18

Annex 1.

OntCork visualised with OWLViz in Protégé

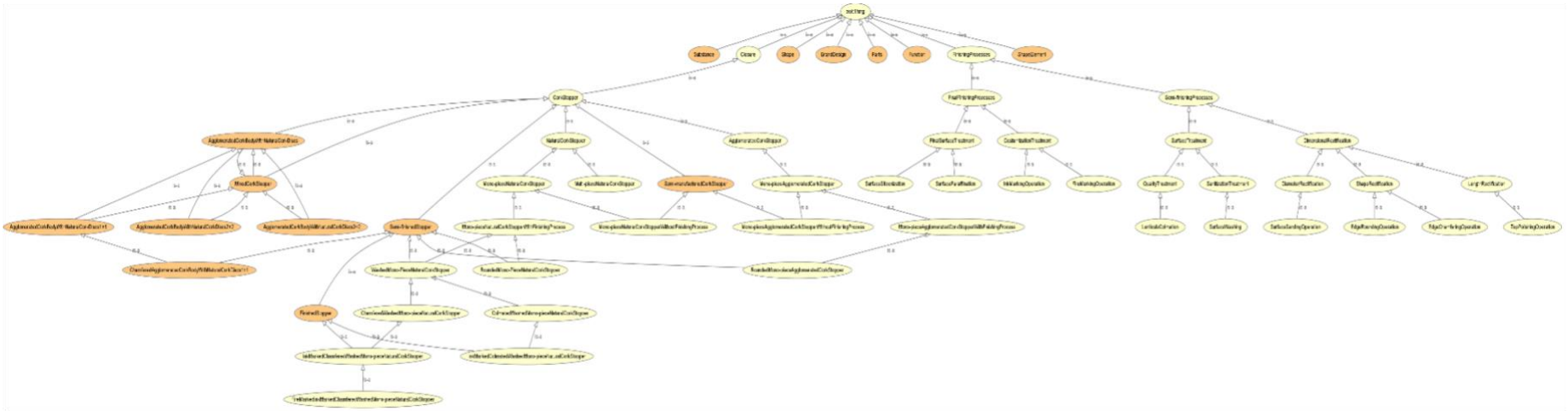

Figure 1. Inferred hierarchy, where the concepts are classified by the reasoner according to additional information, such as their stage of completion, provided by the logical constructs in OWL2.