ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE

CORSO DI LAUREA MAGISTRALE IN
SPECIALIZED TRANSLATION – CURRICULUM TRANSLATION & TECHNOLOGY

THE CLARIN THESAURUS

Constructing a Knowledge Organization System for Indexing Content in Social Sciences and Humanities Research Infrastructures

Tesi di laurea magistrale in Terminology

Relatore
Prof. Adriano Ferraresi

Correlatrici
Dott. Francesca Frontini
Prof. Maja Milicevic Petrovic

Presentata da

Lesley Messori

Sessione luglio 2024

Anno Accademico 2023/2024

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Alma Mater Studiorum Università di Bologna

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE

Corso di Laurea magistrale Specialized Translation (classe LM - 94)

TESI DI LAUREA IN

TERMINOLOGY

The CLARIN Thesaurus
Constructing a Knowledge Organization System for Indexing Content in Social Sciences and Humanities Research Infrastructures

CANDIDDATA

Lesley Messori

RELATORE

Adriano Ferraresi

CORRELATRICI

Francesca Frontini

Maja Milicevic Petrovic

Anno Accademico 2023/2024

Primo Appello

Table of contents

Abbreviation index

Abbreviation	Full form
SKOS	Simple Knowledge Organization System
KOS	Knowledge Organization System
MDS	SSHOC Multilingual Data Stewardship
MM	SSHOC Multilingual Metadata
LRT	Language Resources and Technologies
OWL	Web Ontology Language
RDF	Resource Description Framework
RDFS	RDF Schema
W3C	World Wide Web Consortium
TEI	Text Encoding Initiative
SSHOC	Social Sciences and Humanities Open Cloud

1. Introduction

1.1. Dissertation overview

The present dissertation is the result of a project in collaboration with CLARIN, the research infrastructure for language as social and cultural data, which revolved around the creation of a thesaurus related to the infrastructure and the services and tools it provides.

The Common Language Resources and Technology Infrastructure, commonly known as CLARIN, is a digital infrastructure that operates under the governance and coordination of an ERIC, which stands for European Research Infrastructure Consortium. The primary objective of CLARIN is to facilitate access and dissemination of a great variety of language data, services, and tools for language-based research, especially in the fields of humanities and social sciences (Fišer, Witt, 2022). CLARIN is a distributed infrastructure with several participating centres across Europe and, having Open Science and FAIR data principles as core values, all the tools and data are interoperable and reusable.

The centres are classified according to the kind of expertise and services they offer. B-centres, also known as Service Providing Centres, primarily function as technical centres, granting access to resources; C-Centres specialise exclusively in metadata provision; K-Centres, or Knowledge Centres, focus on specific areas related to language resources and technologies. Their role is fundamental, as they provide expertise and resources in specific domains, actively contributing to the broader development and accessibility of language-related resources and tools.

The primary goal of this project is the creation of a versatile and flexible resource that standardises the specialised vocabulary and systematically organises the domain knowledge within the CLARIN infrastructure. While the main application is the enhancement of content retrieval through topic annotation on the CLARIN website, the project aims for broader utility. The initial version of the CLARIN Thesaurus, developed as an internship project at CLARIN ERIC, resulted in a bilingual English-Italian flat resource (i.e. not hierarchically structured) that laid the foundation for further development. This work seeks to expand and organise the concepts into a structured thesaurus, enhancing its applicability within the CLARIN infrastructure.

The current document is structured into four chapters. Chapter 1. provides an overview of the CLARIN infrastructure, highlighting the reasons behind the creation of the CLARIN Thesaurus. It also examines existing terminological CLARIN resources that supported the development of the thesaurus, both content-wise and methodologically. Chapter 2. begins with a theoretical background, outlining the basic principles of terminology, including definitions of the key notions of ‘concept’,

‘objects’, and ‘terms’, as well as approaches and methods for constructing terminological resources. The chapter then focuses on Knowledge Organization Systems (KOSs), clarifying their roles and differentiating between various types of systems to clarify why a thesaurus is the most appropriate for this work. The final section of this chapter is dedicated to the primary formats and models for representing terminological resources.

Chapter 3. focuses on the Simple Knowledge Organization System (SKOS), the chosen representation format for the CLARIN Thesaurus. It includes a detailed description of the main properties of SKOS and provides examples of three SKOS terminological resources, whose domains align with the subdomains of the CLARIN Thesaurus.

Lastly, Chapter 4. centres on the development of the CLARIN Thesaurus. It begins by detailing the creation process of the first version of the resource, implemented as an internship project. The chapter then shifts to the systematic hierarchical organisation of all concepts included in the first version of the thesaurus, which were categorised into three groups according to the subdomain they pertain to. The classification process for each group is described, highlighting the main identified subclasses. Finally, the chapter discusses the primary challenges encountered during the process.

1.2. Project objectives

As a digital infrastructure, the primary interface of the entire organisation is the CLARIN.eu[1] website, which hosts all relevant information and resources. The content search function, a key service offered by the infrastructure, plays a crucial role and, given the significantly vast amount of content available, it is essential that it operates as efficiently as possible. To achieve this, the content must be accurately annotated and indexed, ensuring that users can effectively interact with the infrastructure and benefit from its services.

However, as a multi-centre infrastructure, the CLARIN website receives contributions from numerous individuals, resulting in significant variability in term choices. Although some terminological resources have been created within CLARIN (see Section 1.2.), there is no comprehensive, official regulation of the specialised terminology used within the infrastructure, which encompasses a wide range of domains. Organising domain knowledge is crucial for any company, especially for large-scale digital infrastructures like CLARIN. Therefore, a terminological resource that standardises the specialised language used within the infrastructure and organises domain knowledge can be particularly beneficial for CLARIN.

[1] CLARIN website: https://www.clarin.eu/

The current work is a report on the development process of a thesaurus that encompasses all relevant concepts regarding the CLARIN infrastructure and the services it provides. As detailed in Section 2.7.3., thesauri, like other terminological artefacts, are used predominantly for vocabulary control and information retrieval. Therefore, the primary application of the CLARIN Thesaurus is the topic annotation of the contents on the website, enabling users to retrieve desired information more easily and rapidly. The thesaurus currently exists as a bilingual English-Italian resource, with plans to localise it into other languages, given the pan-European nature of the CLARIN community. In fact, the content search function is currently available only in English. Implementing a multilingual thesaurus could enable users to search for content in their preferred language, thus enhancing even further the efficiency of content retrieval. Beyond topic annotation for content retrieval, the CLARIN thesaurus can be regarded as a versatile resource that can have multiple applications. For example, it represents an opportunity for national consortia and K-Centres to provide users with definitions about CLARIN's services and tools. Moreover, such a terminological resource can serve translation purposes by acting as a term base for translating and localising the CLARIN ERIC website into other languages.

2. The Basics of Terminology and Its Systematic Management

2.1. Chapter overview

The current chapter focuses on the foundation principles in the field of terminology and the methodologies for constructing terminological resources. The chapter is organised as follows: Section 2.2 covers various foundational aspects of terminology. It begins by defining the objectives of terminology as a multidisciplinary field of study and provides a clear distinction between the fundamental notions of concept, object, and term. The section also introduces key approaches in terminology, namely the onomasiological approach, which contrasts with the semasiological approach typical of lexicography, and the distinction between systematic and ad-hoc terminological work. Additionally, it highlights the importance of text corpora and automatic term extraction in performing terminological tasks, and offers an overview of the structure of a terminological record and the types of information it can contain.

In section 2.3. onward, the focus shifts to Knowledge Organization Systems (KOS), which are frameworks developed to organise and represent knowledge domains, employed across all knowledge domains for different applications, all related to indexing and content retrieval. Section 2.3.1. delves into semantic relations, which are crucial for classifying KOSs and representing concepts within a given domain. Section 2.3.2. provides an overview of the different existing types of KOSs. This is followed by section 2.3.3., which addresses the terminological issues within the realm of KOSs, aiming to clarify the terms used to denote various KOS types. Lastly, Section 2.4. describes the main formats and models used to represent terminological resources.

2.2. Principles of terminology

Cabré (1999) explains how the word ‘terminology’ has three different meanings: the foundational principles and concepts that govern the study of terminology, the guidelines and practices for conducting terminological work, and the collection of terms specific to a particular field. For this reason, Cabré (1999: 32) defines terminology as “an interdisciplinary field of inquiry whose prime object of study are the specialised words occurring in natural language which belong to specific domains of usage”. This definition emphasises the interdisciplinary nature of terminology, as its theoretical principles align with those of linguistics, logic, ontology, and computer science. All these disciplines share a focus on the systematisation of concepts and the organisation of specialised knowledge (Cabrè, 1999). On the other hand, terminology can be viewed as a branch of linguistics, as terms constitute a subset of a language's lexicon. Consequently, it falls within the realm of applied

linguistics, focusing on the practical application of specialised vocabulary. The language employed to describe the knowledge and information of a specific field is known as 'special language' or 'language for specific purposes' (LSP) and is defined as "natural language used in interactions among domain experts, characterised by the use of specific linguistic expressions and communication methods" (ISO 1087:2019, p. 2).

The concept became the focal point in terminology thanks to Eugen Wüster, who is regarded as the father of modern terminology. He theorised that terminological resources should be systematically organised, establishing relationships between concepts to clarify the structure of a knowledge domain (Wüster, 1931). To understand the logic of terminology is fundamental to address the distinction among the three basic notions of *concept*, *term* and *object*. Concepts are defined as "units of knowledge created by unique combinations of characteristics" (ISO 1087-1 2000, p.3.). A concept is intended as the mental representation of an object, real or abstract, and its unique characteristics (Lockinger, Kockaert & Budin, 2015). The process of combining all relevant characteristics of an object to form a concept is called conceptualisation and is the fundamental building block in modern terminology. The process of constructing a concept starts from the identification of a particular object and its core characteristics to define a mental category that describes all objects, abstract or real, that possess those same characteristics. This entails that a concept does not refer to a singular object, but instead refers to a class of objects (Magris et al. 2002). A term is a graphical sign employed to designate the concept and represent it linguistically (Depecker, 2015). As Kageura (2015:47) explains, terms differ from general lexical items in several ways. Firstly, in specialised vocabulary, the meanings of terms are tightly linked to the specific domain or field they belong to, whereas words typically have more universal meanings that do not vary significantly with context. Moreover, in general language there are both content-bearing words, meaning that they refer to an object or entity, and functional elements, like prepositions and conjunctions, while specialised terms are exclusively content-bearing words. Morphologically, terms are primarily nouns, but they can also include verbs and adjectives. Another distinction is that general words are more likely to have synonyms and polysemes. In contrast, specialised vocabularies aim to minimise ambiguity and ensure clarity, resulting in fewer synonyms and polysemous terms. Moreover, most specialised terms are complex, composed of multiple words, whereas simple terms (single-word terms) are less common. According to Nakagawa and Mori (1998, 2002) around 85% of candidate terms extracted for terminological tasks are composed of two or more words. Jacquemin (1999) explains that this happens because multi-word terms carry more semantic richness and specificity compared to single-word terms. Single-word terms tend to be overly polysemous and generic, whereas multi-word terms represent more precise concepts within a domain.

To understand the terminological process, it is relevant to address two important notions of linguistics, onomasiology and semasiology. Semasiology, centres on the meanings of words and how these meanings evolve within a language over time. It examines issues such as polysemy, semantic shifts, and synonymy, making it fundamental to the lexicographical approach of dictionary compilation (Cabré, 1999). Lexicographers start from a list of words, which represents the entries in the dictionary, and describes them as thoroughly as possible, including all the different meanings they can have, across contexts and domains. On the other hand, onomasiology is concerned with concepts or ideas and how they are expressed linguistically (Santos and Costa, 2015). It starts with the definition of a concept and explores the various words or expressions used to denote that concept. For this reason, the onomasiological approach is inherent to terminology, where the focus is to select the right term to define a certain object in a specific context of domain. Terminologists start from a list of concepts pertaining to a specific domain, ensuring that all concepts are logically and ontologically related. These concepts are defined strictly within that domain, and each is assigned a designation or term. When multiple designations exist, one is prioritised based on its usage among specialists in the field.

Cabré (1999) demonstrates that each terminological task, depending on its objectives and domain-specific characteristics, necessitates a different type of terminological search. These searches can be classified based on two criteria that help define the specific approach and methodology required for a given terminological task: the number of languages involved—whether it is a monolingual or multilingual search—and the nature of the search—whether it is systematic or ad-hoc. A search is referred to as 'systematic', or 'subject-field driven', when the terminologist takes into consideration an entire knowledge domain with the goal of creating a terminological resource that represents the specialised language used in that subject area. This involves thoroughly identifying the main concepts, assigning definitions to them, and organising them logically and structurally within one or more conceptual systems (Wright and Wright, 1997). On the other hand, 'ad-hoc', or 'text-driven' terminological search, is conducted when the objective is not to define the entire domain but rather to address specific terminological issues. This approach is particularly relevant for translators who often work with strict deadlines and a variety of subjects, making it impractical to conduct systematic research for every task (Wright and Budin, 1997). Based on the texts they have available, translators create a list of specialised terms and define them to find the corresponding term in another language appropriate for that text.

Corpora, defined as "large collections of authentic texts that have been gathered in electronic form according to a specific set of criteria" (Bowker and Pearson 2002:17), play a crucial role in terminological analysis. Corpora represents a tool that provides dependable contexts that facilitate a

thorough comprehension of terms and the acquisition of domain-specific knowledge. By examining corpora, terminologists can gain deeper insights into the usage and meaning of terms within particular fields, enhancing their understanding and ensuring accurate terminology management. The fact that corpora are in digital form significantly eases the job of terminologists, who are enabled to carry out a linguistic and terminological analysis using specific computational tools that considerably speed up the process.

To build a corpus, the first step is to clearly define the domain under investigation. Once this is established, some of the criteria for selecting the texts to be included in the corpus may vary depending on the research purposes, while others are fundamental. The first is representativeness, to ensure that the collected data accurately reflects the field under analysis. Another crucial aspect is the reliability of the texts, which should be authored by highly reputable individuals (Cabré, 1999). The coverage of the domain is also important, meaning that all significant aspects of the domain should be addressed in the selected texts. The texts should be up-to-date, to reflect the current language used in the field, and should also be written in the target language of the corpus, rather than being translated texts.

Once the corpus is ready, the next step of the terminological workflow is the extraction of candidate terms. Term extraction is a fundamental task that aims to identify the specialised vocabulary of a given domain. Traditionally, it was carried out manually by a terminologist, who would create a list of potential candidate terms after an extensive exploration of the domain and consultation with a field expert. Since the early 1990s, Automatic Term Extraction (ATE) has emerged as a well-established research domain within Natural Language Processing (NLP) and Information Retrieval (IR). ATE aims to alleviate the time-consuming task of manually searching for and selecting terms by automatically identifying candidate terms (Heylen & De Hertog, 2015). It relies on the computational analysis of corpora, providing a more objective analysis of the terms used and their contexts, compared to the subjective judgments of a terminologist or expert, who may have inherent biases. It is important to underline that, despite ATE significantly simplifying the task, it does not fully replace manual procedures. Supervision by a terminologist and validation by a domain expert are still essential to ensure precision.

ATE is carried out using specific tools that can be either commercial or free, web-based or desktop-based. An example is Sketch Engine[2], an online platform that enables users to either build their own corpora or utilise the pre-existing ones available within the system. It offers several features

[2] Sketch Engine: https://www.sketchengine.eu/

to carry out different tasks, including automatic term extraction. The common ground of all these tools is that they compare the vocabulary of special-purpose corpus, with the one of a general reference corpus. The former, also known as focus corpus, is built considering only certain linguistic aspects or a particular domain, while the latter contain general texts with the aim to be as representative as possible of a language (Bowker and Pearson, 2002).

Term extraction is a critical aspect of terminology management, as its output is necessary for carrying out other tasks, depending on the intended use of the candidate terms list. Thurmair (2003) identifies three practical applications: in terminography, applied branch of terminology defined as "terminology work aimed at creating and maintaining terminology resources" (ISO 1087:2019, p.13), the list of candidate terms serves as the input for creating a glossary or database for a specific domain; for **t**ranslation support, where the list of extracted terms is used as an ad-hoc glossary to address precise terminological needs and ensure consistency throughout the translation project; in Information Retrieval, where the candidate term list forms the basis for indexing a document collection, facilitating the retrieval of domain-specific topics for the user.

After the extraction of candidate terms and the validation by domain experts, these have to be organised in relation to one another. There are several types of relationships, among which the main are associative, hierarchical and equivalent. These relationships are addressed in section 2.7.1.

After the identification of all the candidate terms to be conceptualised, all the necessary information can be reported in a terminological record. Cabré (1999) defines a terminological record as "a structured guide that allows us to assign information about a term in an ordered fashion." Terminologists have to select which type of information to include and select the fields that will populate the record, according to the specific needs of the target users of the terminological work and the intended application for the resource. (Cabré 1999). Moreover, different tools or formats used to represent the terminological entries may support different fields.

Drewer e Schmitz (2017) distinguish the most used fields according to the kind of data they allow as input and the degree of liberty they allow to the user. The open fields, such as definitions and contexts, can be filled with any text string, whereas closed fields, like grammatical annotations, restrict the input to a predefined set of options. (Magris et al. 2002). Not every entry has to have all the fields, in fact most of them are optional, while the mandatory ones generally are – beyond the entry label and any alternative labels, such as synonyms, short forms, acronyms – definition, source, context, and equivalents in other languages in the case of multilingual resource (Cabré, 1999). The definition describes a concept as it is conventionally understood within a specific specialised domain, reflecting its representation in texts and the usage by the community of users in that domain (Lockinger, Kockaert & Budin, 2015). Another useful and common open field is the 'notes' field that

can include different kinds of information, such as additional explanations that help to disambiguate the term from others, or any information that cannot be placed in any other field.

2.3. Knowledge Organization Systems

Considering that the aim of the current work is the implementation of a terminological resource that organises the knowledge of the CLARIN infrastructure, it is necessary to dedicate a section to the field of Knowledge Organization Systems (KOSs), shedding light on their purposes and structures.

Knowledge Organization System (KOS) is a generic term that refers to a broad spectrum of schemes designed to organise information and facilitate knowledge management. These systems are developed within the field of Knowledge Organization (KO), which is considered a subfield of library and information science (LIS).

These artefacts play a crucial role in arranging materials for retrieval and managing collections. Acting as a vital link between users' information needs and the resources within a collection, a KOS enables users to identify relevant objects even without prior knowledge of their existence. Whether through browsing or direct searching, on a website or using a site search engine, a KOS guides users through a discovery process. KOSs also allow KOS builders to address questions about the scope of a collection and identify gaps that need to be filled.

The link between terminology and KO is intrinsic, as both disciplines revolve around the systematic organisation of concepts within a knowledge domain. Terminology provides the foundational building blocks for constructing a KOS by establishing standardised, well-defined terms and relationships. In turn, KOS utilises these terminologies to structure, organise, and manage knowledge, making it accessible and useful for various purposes. Therefore, the combined efforts of terminology and KO are essential for enhancing the precision, clarity, and reusability of information in any domain.

2.3.1. Semantic relations

Before discussing the various types of Knowledge Organization Systems and their main characteristics, it is important to address the topic of semantic relations. These relations are crucial as they represent one of the criteria for the differentiation of KOSs types.

Semantic relations are defined as associations between the meanings of words (Miller et al., 1990). A key characteristic of these relations is their reciprocity, which can be asymmetric, where the relationship differs depending on the direction, or symmetric, where the relationship is identical in both directions. Although there are many types of semantic relationships, they can be broadly grouped into three main categories: equivalence, hierarchical, and associative relationships.

Equivalence refers to the relationship that exists between terms that denote the same concept.

Equivalence relationships are always symmetric. Generally speaking, all terms linked by an equivalence relationship are either true synonyms or lexical variants of the same concepts. However, instances where two terms have the exact same meaning in every context, allowing for their substitution without altering the sentence's meaning, are very rare. Thus, synonymy is generally considered relative to context. Therefore, terms that have the same meaning and usage in a wide range of contexts are considered synonyms (Harpring, 2010). For practical purposes, lexical variants are considered synonyms, despite their technical distinctions. Lexical variants are different word forms of the same term, whereas synonyms are generally different terms for the same concept. Lexical variants include spelling variations, abbreviations, and acronyms. For example, in the CLARIN Thesaurus, 'POS tagging' is the preferred label for the concept defined as "a type of tagging in which each word in a text is assigned its appropriate morphosyntactic category." The alternate labels are 'Part-of-Speech tagging' and 'morphosyntactic tagging.' The former is a lexical variant, expanding the acronym, while the latter is a synonym, representing a different term for the same concept.

Harpring (2010) explains how in structured KOSs, equivalence relationships should be established only between terms that are true synonyms to ensure accuracy and precision in indexing and retrieval. However, in resources aimed at retrieval, terms and names with near synonymy (or quasi-synonyms), or similar meanings, may be treated as equivalent to broaden the search results. For example, in the Loterre Open Science Thesaurus (see section 3.10.1.), the terms 'interoperable', 'interoperability', and 'semantic interoperability' are all considered alternate labels of the same concept. Their meaning is not completely identical, and they cannot be used interchangeably in every context, but they are considered equivalent for retrieval purposes. In Knowledge Organization Systems, when multiple synonyms are listed for a concept, only one is designated as the preferred term, also referred to as the descriptor (Harpring, 2010), while the others are regarded as alternates (alternate descriptors). Using the Loterre example, 'interoperable' is identified as the preferred term, while the remaining two are considered alternate labels. When constructing a KOS, creators must set criteria for selecting the preferred term, ensuring consistent application across the resource. Although the primary criterion is choosing the term most commonly used by the majority of users, other factors may also influence the selection. For example, if a specific KOS favours British English spelling, the preferred labels should consistently adhere to British English conventions.

Hierarchical relationships are structured around levels of superordination and subordination, where the superordinate term represents a broader category or whole, and the subordinate terms refer to its members or parts (Zeng, 2008). As discussed in Section 2.7.2, these relationships are what distinguish the simpler forms of KOSs from more structured and complex ones. There are mainly three types of hierarchical relationships, depending on the nature of the relation, namely

genus/species, whole/part, and instance relationships.

The genus/species relation, also known as hyponymy/hypernymy or the “IsA” relation, is transitive and asymmetrical. This semantic relationship describes the connection between two or more concepts so that the meaning of one term encompasses the meaning of another term or terms. In KOSs this relationship is the most prevalent type of hierarchical relation, due to its applicability across a wide range of domains. In Information Retrieval, these are called inheritance systems because a hyponym inherits all the features of the more general concept and adds at least one distinguishing feature, setting it apart from both its superordinate and from any other hyponyms of that superordinate (Touretzky, 1986). For instance, in the initial version of the CLARIN Thesaurus, *K-Centre* and *B-Centre* are categorised as narrower concepts under *CLARIN Centre*. This classification arises from the shared characteristic among all three concepts, namely, their role as components within the CLARIN network. However, *K-Centre* and *B-Centre* distinguish themselves from each other by specialising in particular services, thereby introducing unique features that differentiate them from the broader concept of *CLARIN Centr*e.

The part/whole relationship is also called meronymy/holonymy, or “HasA” relation. This relation applies to instances where one concept is intrinsically contained within another, independent of context. This facilitates the organisation of terms into logical hierarchies, with the overarching concept being recognized as the broader term (Zeng, 2008).

An instance relationship identifies the connection between a general category of things or events, represented by a common noun, and a specific instance of that category, often signified by a proper name. For example, in Loterre the concept *open science project*[3] represent the general category, under which are listed the proper names of several projects pertaining to the category.

Hierarchical relations are not exclusive, meaning that a concept can belong to more than one broader concept, thus constituting a polyhierarchy.

Associative relationships are established when two entities are neither hierarchical nor equivalent, but are conceptually close, so that the connection that exists between them should be explicated (Zeng, 2008). The standard type of associative relationship is denoted as "related to", although some KOSs employ more specific descriptors. The nature and application of these relationships can differ across vocabularies, depending on the terms' characteristics and their intended use in retrieval systems. Every KOSs should define and explicate the types of associative relationships it employs. This type of relationship is primarily used to differentiate terms that are similar in meaning

[3] Concept ‘open science project’ http://data.loterre.fr/ark:/67375/TSO-H93WKMMJ-B

but not identical, thus avoiding potential confusion for users. Generally, associative relations are established between concepts belonging to different hierarchies, as the connection between concepts sharing the same broader concepts is implicit. However, associative relationships can also link sibling concepts—terms that share the same parent concept—when there is some degree of overlap between their meanings.

2.3.2. Classification of KOSs

KOSs are developed for a wide range of domains and have diverse applications, leading to variations in their structures and attributes. The classification of the different types of KOSs is not clear-cut, as they were created in various contexts and time periods, for different purposes, and with differing theoretical and methodological approaches. Mazzocchi (2018) provides a comprehensive review of the criteria used for categorization, including purpose, content, and structure, along with the different proposed classifications. However, the current work focuses on outlining the different types of KOSs to clarify why a thesaurus is the most suitable one for the objectives and applications of a terminological resource regarding the CLARIN infrastructure. Therefore, only one of the proposed classification will be reported. One widely accepted classification is Hodge's (2000), which categorises KOSs based on characteristics such as structure, complexity, relationships, and historical function, resulting in three broad categories: lists, classifications and categories, and relationship models.

Lists are the simplest form of KOS, consisting of a linear collection of terms and their definitions, with no relation established among them. In computer applications, lists are commonly referred to as "flat files" because they lack deep organisation or complex structure. The attributes within lists can consist of simple values or extensive descriptions, and the order of the items typically does not hold intrinsic meaning, often following a logical order, such as numerical or alphabetical, to facilitate retrieval (Pieterse and Kourie, 2014). The most common types of KOSs that fall under the list category are:

- Pick lists (or simply lists): restricted collections of terms organised in a particular sequential manner, such as alphabetical, chronological, or numerical order (Mazzocchi, 2018);
- Dictionaries: alphabetical lists of terms with their definitions, which typically encompass additional details such as spelling, morphology, origin, and variant senses for each term. While some dictionaries may incorporate cross-referencing between related entries, these references are not considered integral components of the dictionary's structure and are therefore are still classified as lists (Pieterse and Kourie, 2014);
- Glossaries: alphabetical compilations of terms regarding a particular domain, along with their corresponding definitions (Mazzocchi, 2018);

- Synonym rings: sets of terms that, for information retrieval purposes, are considered equivalent, even though they are not true synonyms. In synonym rings there is no distinction between preferred and non-preferred terms, as they allow to broaden the search and retrieve more information. For this reason the alternative terms are not displayed to the user. (Harping, 2010);
- Authority files (or name authority lists): lists of terms used for controlling variant names for an item, where one term is designated as the preferred one. The primary feature of these lists is that non-preferred terms are indicated to the user and function as cross-references to direct them to the preferred term (Mazzocchi, 2018);
- Directories: compilations of named entities, such as people, places or institutions, accompanied by their respective contact details (Mazzocchi, 2018);
- Gazetteers: a geographical compendium or reference tool, often utilised alongside maps or atlases. Its contents encompass various aspects of geographical composition, social statistics, and physical characteristics pertaining to a country, region, or continent. This information is usually organised into thematic categories, with entries listed alphabetically for easy reference.

The second group is classification and categories, where the common characteristic is the presence of hierarchical relationships among the elements of the categorization. The main types of KOSs that fall under this category are:

- Subject headings: these are controlled vocabularies comprising terms that represent the subjects of items within a collection, together with rules for combining these terms into compound headings (Mazzocchi, 2018). An illustrative example of this is the *Medical Subject Headings (MeSH)*[4], which is a hierarchically-organised vocabulary produced by the National Library of Medicine for indexing, cataloguing, and searching of biomedical and health-related information.
- Taxonomies: hierarchically organised collections that contain items and their attributes. (Pieterse and Kourie, 2014). The term has been used since the 1700s to refer to the systematic organisation and naming of living organisms. In a taxonomy, elements within a domain are grouped into categories and sub-categories, which can be several levels deep, forming hypernyms and hyponyms relations.

[4] Medical Subject Headings (MeSH): https://www.nlm.nih.gov/mesh/meshhome.html

- Classification schemes: these are hierarchical and faceted organisational structures comprising numerical or alphabetical notations. They serve the purpose of representing broad topics and are typically designed as universal systems, encompassing all fields of knowledge. Among the most renowned classification schemes is the *Dewey Decimal System (DDC)*, published for the first time in 1876.

Finally, the third category comprises relationship models, recognized for their complexity and highly structured nature, which focus on the interconnection between terms and concepts. The typologies of KOSs that adhere to these characteristics are:

- Thesauri: these are controlled and structured vocabularies that exhibit hierarchical, associative, and equivalence relations among concepts within a specific domain (Mazzocchi, 2018).
- Semantic networks: these systems represent terms or concepts as nodes within a network, with various types of relationships connecting them. They are more elaborate than thesauri in defining categories or semantic types and the relations between them. For instance, the UMLS (Unified Medical Language System)[5] Semantic Network, which deals with biomedical terminology, encompasses 135 semantic types and 54 relations.
- Ontologies: these are typically described as formal, explicit specifications of a shared conceptualization (Gruber 1993). They often comprise intricate relations between entities and incorporate rules and axioms to facilitate logical reasoning. They also offer properties and instances, and serve as conceptual vocabularies, enabling information retrieval, knowledge reuse, and the automatic derivation of new knowledge.

Categories of KOSs	Common category characteristics	Types of KOSs
Lists	Linear collections of terms with no structural relations	Pick lists Dictionaries Glossaries Synonym rings Authority files Directories Gazetteers
Classification and categories	Hierarchically structured systems	Subject headings Taxonomies

[5] Unified Medical Language System (UMLS): https://www.nlm.nih.gov/research/umls/index.html

		Classification schemes
Relationship models	Structured systems with complex semantic relations among the concepts	Thesauri Semantic networks Ontologies

Table 1. Summary of KOSs types

2.3.3. Terminological issues

The existence of various classifications of Knowledge Organization Systems (KOSs), each grounded upon different criteria, poses a notable challenge concerning the terminology used to describe them. Mazzocchi (2018) highlights that several authors have emphasised inconsistencies in the literature regarding the terminology for KOSs. It is paradoxical that there is "a serious lack of vocabulary control in the literature on controlled vocabulary" (Weinberg, 1998). Indeed, numerous terms referring to KOS types lack precise definitions, resulting in ambiguity and overlapping usage across diverse communities of practitioners and contexts.

The first term that is used ambiguously is ‘controlled vocabulary’, which is sometimes used as an umbrella term for several, if not all, types of KOSs. For example, Harpring (2010) uses this term basically as a synonym of “Knowledge Organization systems” and defines controlled vocabularies as

> organised arrangements of words and phrases used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms and has a defined scope or describes a specific domain. (12)

Harpring (2010) further elaborates on the different types of controlled vocabularies and their characteristics, covering all the types mentioned in Section 2.7.2. Hedden (2008) also treats controlled vocabularies as a synonym of KOS, but she poses a distinction between simple controlled vocabularies, such as lists, and taxonomies and thesauri, due to their complex structure. In contrast, Soergel (2009b) views controlled vocabulary as just one type of KOS, defining it specifically as a subtype of an authority file.

Another term that creates confusion is ‘taxonomy’. Gilchrist (2003:11) highlighted that the term ‘taxonomy’ is employed with at least five distinct, yet overlapping, meanings: in the context of web directories, it denotes website dropdown menus facilitating navigation to further levels of content; in the realm of taxonomies supporting automatic indexing, it signifies algorithms comprising sets of words, phrases, synonyms, and syntactic variations utilised for extracting index terms automatically; in the context of taxonomies generated through automatic categorization, it relates to

software packages analysing texts and generating categories for document classification automatically; it also encompasses front-end filters, where taxonomies are either created or imported for use in query formulation; lastly, in the domain of corporate taxonomies, it represents resources developed within a specific enterprise to assist staff and users in browsing and searching portals. Gilchrist's analysis reveals that all five meanings associated with the term ‘taxonomy’ are interconnected. Instead of displaying features common to all, this term refers to a set of related items that are best understood through Wittgenstein's (1953) notion of family resemblance. In other words, they share similarities in various ways, as it occurs with members of a family.

The same happens with the term ‘thesaurus’. Spärck-Jones (1992) outlines three primary meanings associated with the term. The first is to denote a vocabulary reference work to aid writing, leading to the denomination of ‘Vocabulary Reference Thesaurus’ (VR thesaurus). The second meaning applies to the Library and Information Science (LIS) field, where thesaurus is a tool for vocabulary control, ensuring consistency in item descriptions and facilitating retrieval. Here, the emphasis lies not on defining concepts individually but on delineating the relationships between concepts and disambiguating them. In the LIS field, this type of resource is known as an information retrieval thesaurus (IR thesaurus). The third meaning is rooted in its application within the field of Artificial Intelligence, where it denotes a repository of words and phrases commonly used as a resource for natural language processing, hence referred to as a Natural Language Processing thesaurus (NLP thesaurus). Additionally, Mazzocchi (2018) discusses metathesauri, which seek to integrate existing thesauri and vocabularies. He also addresses automatically constructed thesauri, where relationships are established by computer algorithms, and which typically exhibit a less structured semantic organisation compared to standard IR thesauri.

Another factor contributing to ambiguity is the blurred distinction between thesauri and taxonomies. Often, the boundaries between these two concepts are not clearly delineated, leading to their interchangeable usage. Pieterse and Kourie (2014) elucidate that the distinction between thesauri and taxonomies lies in the kind of relations they support. Taxonomies typically utilise hierarchical relations exclusively, whereas thesauri accommodate various other types of relations, although the granularity may vary from depending on the application. The types of semantic relations that can be established between elements in a thesaurus can be categorised into four main types: equivalence, hierarchical, associative and contrast. Equivalence, hierarchical, and associative relations are commonly associated with information retrieval (IR) thesauri (Tudhope & Binding, 2008), while Vocabulary Reference (VR) thesauri primarily utilise equivalence and contrast relations. Natural Language Processing (NLP) thesauri typically incorporate all of these types of relations and require a finer granularity in distinguishing between different relation types compared to other types of

thesauri.

While ambiguity exists across the labels used to denote various kinds of knowledge organization systems (KOSs), the term ‘ontology’ stands out as particularly confusing. Ontologies are commonly understood as the "philosophical study of being in general, or of what applies neutrally to everything that is real" (Encyclopaedia Britannica). In simpler terms, ontologies strive to delineate and represent all facets of human knowledge, aiming to capture and organise information about entities and their relationships within a specific domain. Since the 20th century, the prevailing approach to ontology has centred on both logic and linguistic methods. These methodologies rely on theories of meaning and reference, which are applied to artificial logical languages or natural languages, in order to discern the types of entities that exist. In an attempt to incorporate the perspective of concept ontology as used in LIS, Pieterse and Kourie (2014:223) define it as “an electronically stored collection that comprises a thesaurus combined with a set of inference rules.” In comparison to other types of Knowledge Organization Systems (KOS), ontologies can be viewed as an extension of thesauri. While thesauri primarily organise concepts with few kinds of semantic relationships, ontologies incorporate attributes and more complex relationships in a formalised structure. The distinguishing characteristics of ontologies include the necessity of a formalism as the basis for representation and the integration of inference rules. These rules are essential for encoding information to enable manipulation and interpretation by computer programs. Furthermore, inference rules facilitate greater semantic expressiveness, enabling more detailed information about concepts, deeper hierarchical levels, and richer relationships between concepts.

Since ontologies are inherently more structured and machine-readable compared to thesauri, practitioners frequently strive to transform existing thesauri into formal ontologies to facilitate automatic reasoning. This transformation requires the formalisation of data using dedicated standards and the incorporation of inference rules. To facilitate this process, various technologies have been developed within the Semantic Web framework for encoding ontological information and promoting interoperability. These include the Resource Description Framework (RDF) (Klyne and Carroll, 2004), the Web Ontology Language (OWL) (McGuinness and van Harmelen, 2004), and the Simple Knowledge Organization System (SKOS) (Miles and Bechhofer, 2009). These frameworks are addressed in Section 2.8. and Chapter 3.

This overview of different types of Knowledge Organization Systems (KOS) and their attributes aids in identifying the reasons why a thesaurus is the most suitable type for the intended applications of a terminological resource related to CLARIN. As outlined in the project objectives, the primary goal is to annotate content on the main CLARIN website and other CLARIN content to facilitate retrieval and enhance the user experience, particularly within the CLARIN community.

Given CLARIN's role as a research infrastructure, a resource aimed at organising the enterprise's knowledge for user benefit could be considered a corporate taxonomy. However, considering the structural disparities between taxonomies and thesauri—specifically, that taxonomies solely employ hierarchical relations while thesauri support various types of relations—it appears more appropriate to designate the CLARIN terminological resource as a 'thesaurus'. This decision allows for a more comprehensive description of concepts and their interconnections.

2.4. Formats and models for representation

There are several formats and standards developed for the implementation of KOSs. Most of them are developed and/or maintained by the World Wide Web Consortium (W3C)[6] the international organisation operative since 1994, dedicated to establishing and maintaining standards and guidelines that influence the development of the World Wide Web (WWW), with a focus on principles such as accessibility, internationalisation, privacy, and security.

Hyvönen (2002) explains how the first generation of the WWW, which surfaced in the early 90s and was based on HTML, facilitated document access and visualisation on the internet, separating presentation from document location. Meanwhile, the second generation, which developed in the late 90s, and relied on XML[7] and other Markup Languages, separated the document structure from its presentation, evolving the purpose of the web from merely a place for users to visualise documents, to a place where to store data, which can be then represented in different ways according to their purpose. As a markup language, XML allows the user to annotate the text to make explicit any information for a computer program. Being a metalanguage, XML provides syntactic guidelines that can be applied across various formats, ensuring that content is both comprehensible to humans and processable by machines. XML is widely used for generating source documents due to its ability to support serialisation—the process of converting objects or data structures into a format suitable for storage or transmission (Roturier, 2019). A key feature of XML is its extensibility. Unlike markup languages with fixed tags, XML allows users to create custom tags while adhering to a standardised syntax. This flexibility ensures that XML documents can be seamlessly exchanged between different programs without losing any data.

There are numerous standards and formats based on XML, including the TEI guidelines (TEI

[6]W3C website: https://www.w3.org/

[7] Extensible Markup Language (XML), W3C: https://www.w3.org/XML/

Consortium). The Text Encoding Initiative (TEI)[8] is a consortium focused on developing and maintaining standards for representing texts in digital form. Using XML and Unicode, TEI guidelines provide a flexible framework for encoding textual information, ensuring data integrity and interoperability across various software and systems. This facilitates seamless exchange and reuse of electronic texts, especially in the humanities. The guidelines are adaptable to any natural language, period, literary genre, or text type, allowing customization for different purposes and enhancing their applicability in diverse digital humanities projects.

Another notable standard is Term Base eXchange (TBX), an open format designed for the interchange of terminological data. Initially created by the Localization Industry Standards Association (LISA), TBX is currently managed by the International Organization for Standardization (ISO) under the designation ISO 30042. TBX is a concept-oriented data model, i.e. senses (concepts) are considered the organising principle of the terminology database. It allows the management of complex terminological information, such as preferred and alternate labels, definitions, scope notes and other metadata. it also supports multilingual terminological entries. TBX is designed as a flexible framework capable of accommodating a broad spectrum of user-defined requirements and data models. It offers a great variety of data categories that can be selected and combined to form different TBX dialects. These dialects ensure that terminology data can be customised to meet the diverse needs of various applications, industries, or organisations while still adhering to a common framework. Some dialects, such as TBX-Default with 117 data categories and TBX-Basic with 29 data categories, are publicly recommended by ISO 30042 (Reineke, 2014). Other dialects, privately created by users, are not endorsed or recommended by ISO 30021.

Today, the majority of standards employed for the formalisation of KOS are part of the realm of the semantic web, which corresponds to the third generation of the WWW, where also the meaning of documents is separated from their structure (Hyvönen, 2002). The Semantic Web is therefore defined by Berners-Lee et al., (2001:1), founder of the WWW, as "an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." In fact, the initial structure of web pages allowed users to have easy access to a great amount of information, but it posed a limitation: the content was primarily intended for human consumption, making it challenging for computers to grasp the semantics. Consequently, while computers could retrieve information, they lacked the ability to comprehend it. This constraint became apparent especially in the face of more complex queries, demanding increased human

[8] TEI: https://tei-c.org/

intervention. Therefore, the overarching goal of the Semantic Web was to organise the content of web pages in a format beneficial for machine understanding, with the ultimate objective of enhancing the flexibility and automation of computers and web searches.

The same article (Berners-Lee et al., 2001), successively illustrates how, in order for the semantic web to work effectively, a shift in knowledge representation methods and systems was imperative. Traditionally, these systems were predominantly centralised, requiring universal consensus on identical definitions of common concepts. However, managing knowledge in a centralised manner proved impractical, considering the vast quantity of content present on the web. Moreover, traditional knowledge representation systems operated with distinct and limited sets of rules for making inferences about their data. While these systems facilitated data exchange between one another, the exchange of rules between different systems was not feasible. Thus, the Semantic Web aimed to exploit a language capable of expressing both data and the rules governing its interpretation.

In the technological context, the term semantic web is employed to denote "a set of technologies, tools and standards which form the basic building blocks of a system that could support the vision of a Web imbued with meaning." (Matthews, 2005:4). Notably, the Semantic Web is characterised by a layered structure, as illustrated in Figure 2. below, where each layer represents a crucial technological component, necessary for the realisation of a web capable of decoding the meaning of content and linking resources.

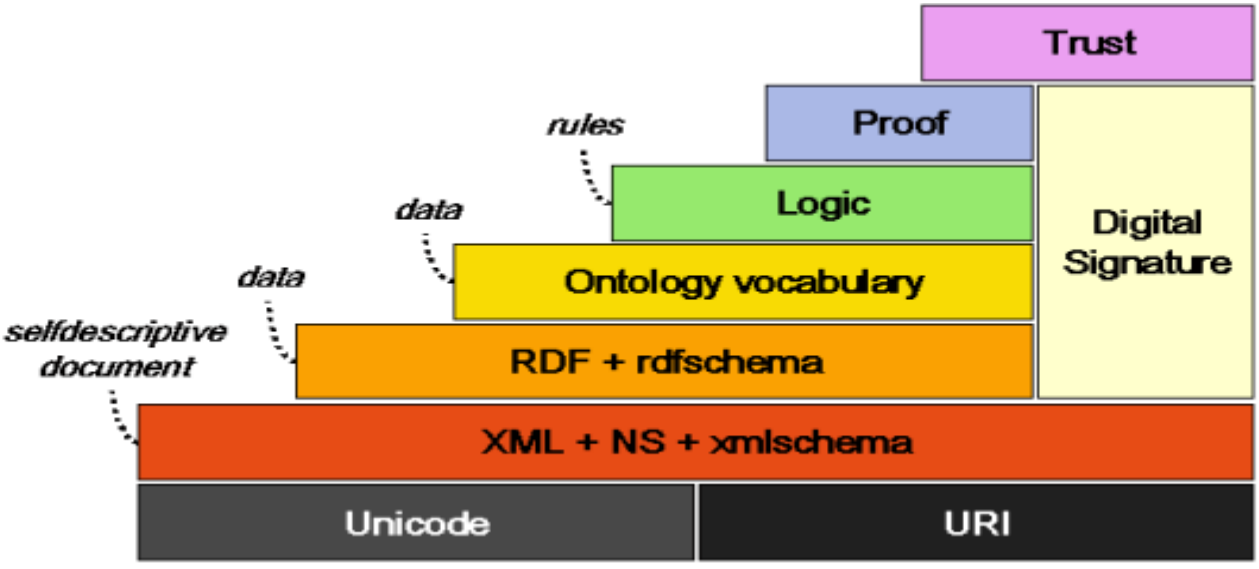

Figure 2. Semantic Web Layers (Koivunen, Miller, 2002)

Unicode and URIS: The foundation of the Semantic Web is built upon Unicode, the standard for computer character representation, and URIs (Uniform Resource Identifier), the standard for the identification of resources (Berners-Lee, 2005). Both these standards play a fundamental role in

ensuring the effective identification and reusability of resources within the Semantic Web framework.

XML: The second layer is constituted by XML and its associated standards, including Namespaces and Schemas. This layer is designed to structure web data and facilitate the integration of Semantic Web definitions with other XML-based standards (Koivunen, Miller, 2002). However, it is important to note that XML, while enabling users to create their own tags to label and structure documents, does not inherently convey the meaning of the data structure itself. In fact, according to Berners-Lee et al. (2001), although XML allows for arbitrary tagging and structuring of documents, it falls short in expressing the semantic meaning of that structure.

RDF & RDFS: The meaning is transmitted thanks to the next layer, composed of the Resource Description Framework (RDF) and RDF Schema (RDFS). RDF is "a standard model for data interchange on the Web" (RDF Working Group, 2014), that encodes sentences in sets of triplets, which predominantly correspond to subject, verb and object. The triplets are written using XML tags and each element is identified through a URI. The RDF triples collectively create webs of information about related entities. Crucially, RDF employs URIs to encode data within a
document, ensuring that concepts extend beyond mere textual representation and are linked to distinct definitions universally accessible on the Web (Berners-Lee et al., 2001).
RDF Schema is an application of RDF and can be considered a primitive ontology language, offering modelling primitives with fixed meanings for describing specific knowledge domains (Grigoris, van Harmelen, 2004). Thanks to RDFS it is possible to define vocabularies, identifiable by URIs (Koivunen, Miller 2002).

Ontology vocabulary: The subsequent stratum in the Semantic Web architecture is embodied by Ontology vocabulary, which provides a more intricate and sophisticated language for delineating specific domains (Grigoris and van Harmelen, 2004). Koivunen and Miller (2002) state how this layer of ontology vocabulary facilitates the progression of vocabularies by enabling the definition of relationships between distinct concepts.

Logic, Proof, and Trust: The upper layers of the structure, namely Logic, Proof, and Trust play a crucial role in implementing and managing the rules employed to represent knowledge. The Logic layer facilitates the formulation of rules, while the Proof layer carries out the execution and, collaborating with the Trust layer, it evaluates whether to place trust in the provided proof or not (Koivunen, Miller, 2002).

Digital Signature: In their analysis of the Semantic Web structure, Koivunen and Miller (2002) provide insights into the standardisation process within W3C working groups. Notably, most of the layers outlined so far, namely RDF, RDF Schema, Ontology vocabulary, Logic, and Proof, have undergone standardisation, supervised and authenticated by the Digital Signature layer,

designed to detect alterations to documents.

The Ontology vocabulary layer consequently fully developed into the Web Ontology Language (OWL), which is defined by the W3C as a "Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things". (OWL Working Group, 2012). The initial version of OWL was developed by the W3C Web Ontology Working Group[9] and published in 2004. In 2009, an extension of the standard, named OWL 2 and curated by the W3C OWL Working Group[10] was published, and its Second Edition, which represents the current version, was released in 2012 (OWL Working Group, 2012).

OWL is a formal language, built upon RDF and RDFS, used to construct complex ontologies that thoroughly describe the relationships between concepts within a particular domain. Compared to RDF, OWL allows a higher level of expressiveness, supporting more complex and detailed descriptions of concepts and their interconnection. Moreover, it supports inference, allowing logical reasoning over the defined data. Another format based on RDF is the Simple Knowledge Organization System (SKOS), a W3C recommendation for representing various KOSs used in indexing and classification. As the name implies, SKOS supports a simpler representation of concepts and their relationships compared to more complex ontological formalisms like OWL. Given that SKOS is the chosen format for implementing the CLARIN Thesaurus, its functionalities and characteristics are discussed in detail in Chapter 3.

Within the semantic web, the role of ontologies has become more and more relevant for modelling and representing domains in a variety of forms. (McCrae et al., 2017). However, the available ontology languages like OWL and RDF(S) fell short in supporting the incorporation of linguistic information, particularly in detailing how ontologies entities, such as properties, classes, and individuals,can be expressed in natural language. The Lemon model (McCrae et al., 2017) was implemented to compensate for this need and has become the primary mechanism for the representation of lexical data. Lemon was further developed in the context of the W3C OntoLex community group, resulting in the OntoLex-Lemon model (Cimiano, et al., 2016). Ontolex-Lemon seeks to bridge the gap by offering a vocabulary that enables the integration of linguistic information into ontologies. This integration specifies how vocabulary elements within these ontologies are expressed in natural languages. For example, it allows the representation of morphological and syntactic properties of lexical entries. The aim is to render ontologies more accessible and interactive

[9] https://www.w3.org/2001/sw/WebOnt/

[10] https://www.w3.org/2007/OWL/wiki/OWL_Working_Group

for human users, and to enhance the compatibility of ontologies with NLP tools (Cimiano, et al., 2016). Moreover, the separation of the ontological and lexical layer facilitates the multilingual adaptation of an ontology, by simply changing the lexicon from one language to another.

Furthermore, in the last decades the terminology community has shown a growing interest in converting terminological data from XML-based formats like TBX into Semantic Web formalisms to adhere to the FAIR principles of reusability and interoperability and to support Open Science. For instance, Reineke (2014) introduced an almost fully automated conversion routine that maps TBX data models to RDF/XML serialisation and then reconverts the RDF document back to TBX without any data loss. Another notable example is the TBX2RDF conversion system (Cimiano et al., 2015; Montiel-Ponsoda et al., 2015), which integrates TBX resources into the Linked Open Data (LOD) framework by converting them into the Ontolex-Lemon model. Additionally, Bellandi et al. (2023) developed an interactive TBX to Ontolex-Lemon converter that allows end users to actively participate in the conversion process.

3. Simple Knowledge Organization System (SKOS)

3.1. History and development of SKOS

The Simple Knowledge Organization System (SKOS) (Miles, Bechhofer, 2009) is a W3C recommendation for the development and representation of a multitude of Knowledge Organization Systems (KOS) for indexing and classification purposes, such as thesauri, taxonomies and other types of controlled vocabularies (See Section 2.7.2.). SKOS employs the Resource Description Framework (RDF) (Klyne and Carroll, 2004) to standardise and enhance interoperability among these systems (Smith, 2022).

In Smith (2022) a brief timeline of the evolution and adoption of SKOS as a W3C recommendation is traced. The first draft of SKOS was released in 2004 (Miles, Rogers, Beckett, 2004), as a product of the Semantic Web Advanced Development (SWAD) team. This effort, initiated in 2001, built upon the groundwork laid by earlier European projects like DESIRE (Development of a European Service for Information on Research and Education), from 1997 to 2000, and LIMBER (Language Independent Metadata Browsing of European Resources) from 1999 to 2001. After its first publication in 2004, SKOS transitioned into a working draft under the guidance of W3C. In 2006, the Semantic Web Deployment working group (SWD) conducted a comprehensive review, culminating in the formal acknowledgment of SKOS as a W3C Recommendation in 2009 (Miles, Bechhofer).

3.2. SKOS rationale

Thomas Baker et al. (2013) illustrate the rationale that stands behind the development of SKOS. The development of Knowledge Organization Systems (KOS) is driven by specific purposes and use cases (See Section 2.7.). Primarily, these systems aim to facilitate the retrieval of objects from indexed collections, leveraging hierarchical and associative relations established among them. This aligns with the capabilities provided by Semantic Web technologies, enabling users to effortlessly reuse data from diverse contexts and establish links between various KOSs. In other words, the authors explain how SKOS represents a solution for the integration of the realm of KOSs into the Semantic Web at a low cost by expressing features common to a wide range of KOS types. This approach proves faster and more cost-effective compared to translating KOSs. This is because KOSs are typically designed as informal structures reflecting human intuitive knowledge. Translating them into the formal languages of RDFS and OWL involves mathematical formalism, defined rules, and reasoning that enforce constraints or generate new knowledge through inference. Thus, SKOS offers a simple,

intuitive conceptual modelling language for creating and sharing new Knowledge Organization Systems (KOSs), serving as a bridging technology between the strict logical formalism of ontology languages such as OWL with the weakly-structured, informal XML-based formats and standards.

For these reasons, SKOS was designed with the principle of maintaining a *minimal ontological commitment* when defining the concepts and the relationships between them. As Thomas Gruber (1995:100) states:

> An ontology should require the minimal ontological commitment sufficient to support the intended knowledge sharing activities. An ontology should make as few claims as possible about the world being modelled, allowing the parties committed to the ontology freedom to specialise and instantiate the ontology as needed. (12)

This principle was adopted to avoid over-commitment in the formalisation of SKOS features and to prevent overlap and interference with OWL and RDF specifications. Consequently, SKOS is crafted to capture the essential, informal semantics most commonly required for KOS purposes, reflecting the thesaurus standards ISO 2788 and ISO 5964, although not adhering to them entirely. Moreover, a key objective was to make SKOS as flexible as possible, enhancing its applicability across a broader spectrum of applications. As a result, the recommendations of SKOS are not viewed as rigid constraints but are rather considered as recommended best practices.

In fact, SKOS is intended to support a diverse range of Knowledge Organization System (See Section 2.7.), such as glossaries, thesauri, taxonomies, each with their specific characteristics. The common ground of all types of controlled vocabularies is the organisation of knowledge through the aggregation of a coherent set of lexical entities, for instance terms, words, headings, captions, etc. As stated in the SKOS Reference (Miles, Bechhofer, 2009: Section 3), these lexical entities are referred to as ‘concepts’:

> A SKOS concept can be viewed as an idea or notion; a unit of thought. However, what constitutes a unit of thought is subjective, and this definition is meant to be suggestive, rather than restrictive.

This broad definition of concept aligns with the flexibility and adaptability that SKOS aims to achieve, allowing users to adopt the definition of concept that best suits the type and scope of the resource.

3.3. SKOS features

As previously mentioned, the focal points in the SKOS model are concepts, each uniquely identified by a URI and labelled with lexical strings in one or more natural languages. Each concept can be annotated with different types of notes, each serving distinct purposes. Furthermore, concepts can be

linked through hierarchical or associative relationships that are not governed by formal ontological rules, thereby creating informal hierarchies (concept schemes) and association networks. Additionally, concepts can be mapped to those in other concept schemes, reinforcing the interconnected nature of the SKOS framework (Smith, 2022).

A concept scheme is a cluster of concepts pertaining to a specific domain and the relations established among them. It is expressed by the `skos:ConceptScheme` class, representing the top-level organisational structure in SKOS. Concept schemes are not mutually exclusive, as a single concept, identified by its URI, can belong to multiple schemes and occupy different positions within their respective hierarchies. It is crucial to emphasise that concepts and concept schemes are distinct entities, each identified by a unique URI. In fact, it is possible that within a concept scheme, the same label is used to refer to both the concept scheme itself and a concept, often the top concept.

Beyond the aggregation of concepts into schemes, SKOS also makes it possible to group them into 'collections' of concepts with the `skos:Collection` class. It is important to highlight that SKOS defines `skos:Collection` as disjoint from `skos:ConceptScheme` and `skos:Concept`. In fact, collections cannot be used in combination with semantic relations to assign the instances a position in the semantic structure of a KOS. Nevertheless, users still have the possibility to determine the order in which instances in a collection are displayed, using the `skos:OrderedCollection` class. A crucial distinction between hierarchical organisation (concept schemes) and collections lies in the instances that function as nodes: whereas in a hierarchy, broader concepts serve as nodes for further branches, in collections, nodes are simply labels for a specific grouping and do not represent real concepts. Frequently, they are not even identified with a URI (Smith, 2022).

In the upcoming sections, each of the features offered by SKOS will be addressed, with a particular focus on aspects relevant to the current project's scope. It is essential to underline that while the coverage of each topic will be concise, the intricacies of the SKOS framework could allow for more in-depth exploration. The information about the functionalities and structure of SKOS is derived from Arthur Smith (2022). This paper provides a comprehensive account of the history and development of SKOS, elucidates the SKOS Recommendations (Miles & Bechhofer, 2009), and offers insights into its potential applications, adoption possibilities, and recommended practices.

3.4. SKOS Labels

As stated above, the fundamental element in SKOS is `skos:Concept`, representing a unit of meaning identified by a URI, corresponding to the subject or object in RDF. For each concept, various labels can be associated, represented by string literals that can be expressed in any language and

identified by language tags consisting of 2-letter ISO language codes, which renders SKOS completely multilingual. Labels consist of the terms identified to denote a given concept.

Three types of labels are possible, one being mandatory and the other two optional: `skos:prefLabel`, representing the primary denomination of the concept and typically being unique and obligatory for each concept; `skos:altLabel`, indicating additional words or phrases synonymous with the preferred label, such as synonyms, acronyms, or spelling variations. The alternate labels, when searched, lead to the same concept and may or may not be displayed to end users; finally, `skos:hiddenLabel`, which shares similarities with alternate labels but is primarily used for misspelt forms, remaining hidden from end users and only employed for searching purposes without being displayed. The choice to have a URI as an identifier stems from the fact that overtime labels could be subject to change, while the concept represented by the URI remains constant and unalterable.

3.5. Semantic relations in SKOS

When addressing semantic relations in SKOS, the properties that fall under this category are `skos:narrower`, `skos:broader` and `skos:related`. In contrast to formal ontologies, in SKOS semantic relations are less strictly and granularly defined, as they primarily serve as an aid for information retrieval. In SKOS when a hierarchical relationship is placed, it is not specified whether it is a genus/species, whole/part, or instance relationship, thus giving the user the flexibility to determine how to employ these properties.

The sole constraints are, for hierarchical relationships, that `skos:broader` and `skos:narrower` form an inverse property pair, describing a bidirectional, non-transitive hierarchical relationship. This means that these properties should only be used to assert direct links between concepts. On the other hand, for associative relationships, the property `skos:related` is symmetric, acknowledging that the association between two concepts is independent of direction. Consequently, hierarchical and associative relationships are declared to be disjointed in SKOS (Baker et al., 2013).

Additionally, SKOS supports polyhierarchy, allowing a narrower concept to have multiple broader concepts. This does not create any disarray because, when a narrower concept is displayed, all the broader concepts are indicated (See Section 3.4.1., Figure 4.). The employment of URIs for identification ensures that a child concept can retain the same meaning and be retrieved independently of its parent concept. (Smith, 2022).

3.5.1. Documentation properties

To provide additional details on the concepts, SKOS offers a variety of documentation properties designed to annotate them with diverse information. The most general is `skos:note`, which has seven subproperties, will be listed below together with their definition as stated in Isaac, Summers (2009):

- `skos:definition`: supplies a complete explanation of the intended meaning of a concept;
- `skos:scopeNote`: supplies some, possibly partial, information about the intended meaning of a concept, especially as an indication of how the use of a concept is limited in indexing practice;
- `skos:example`: supplies an example of the use of the concept's label;
- `skos:historyNote`: describes significant changes to the meaning or the form of a concept:
- `skos:changeNote`: documents fine-grained changes to a concept, for the purposes of administration and maintenance:
- `skos:editorialNote`: supplies information that is an aid to administrative housekeeping, such as reminders of editorial work still to be done, or warnings in the event that future editorial changes might be made:

The last two properties, namely `skos:changeNote` and `skos:editorialNote`, are particularly beneficial for developers and editors of the resource, as indicated by their respective definitions.

3.5.2. Mapping properties

A key feature of SKOS are its mapping properties, which enable the establishment of semantic relations between concepts across different schemes and domains. This capability is made possible by the unique identification of concepts through URIs, ensuring their independent and unambiguous definition regardless of the context. These mapping properties are particularly significant as they enable the reuse of terminological resources and contribute to the standardisation of classification schemes.

The property `skos:exactMatch` is specifically designed for concepts that share an identical meaning. However, given that vocabularies are often maintained by different individuals, it is unlikely for the meanings assigned to concepts to be precisely the same. As a result, to establish exact match relations, it is adequate for the intended meanings to be sufficiently close for interchangeability. In cases where the meanings are not precisely identical, or they do not reach a certain degree of similarity, the property `skos:closeMatch` is provided. To set hierarchical cross-

scheme relations, the properties `skos:narrowMatch` and `skos:broadMatch` properties are utilised, whereas `skos:relatedMatch` can link concepts that share an associative relation.

Types of property	Property functions	SKOS properties

Label properties	Allow the display of the preferred and alternate term do designate a concept	`skos:prefLabel`
		`skos:altLabel`
		`skos:hiddenLabel`

Relation properties	Allow the establishment of hierarchical and associative relations between concepts	`skos:narrower`
		`skos:broader`
		`skos:related`

Documentation properties	Allow the appointment of any additional information that may help to define and disambiguate concepts	`skos:definition`
		`skos:scopeNote`
		`skos:example`
		`skos:historyNote`
		`skos:changeNote`
		`skos:editorialNote`

Mapping properties	Allow the linking and establishment of semantic relations among concepts belonging to other existing resources	`skos:exactMatch`
		`skos:closeMatch`
		`skos:narrowMatch`
		`skos:broadMatch`
		`skos:relatedMatch`

Table 2: Summary table of SKOS properties

3.6. Existing SKOS resources within CLARIN and SSHOC

Although no comprehensive standardisation of the specialised terminology used within the CLARIN infrastructure has been carried out yet, a few existing terminological resources were developed by CLARIN. For instance, there is a glossary,[11] that collects the main acronyms employed within the infrastructure, together with their expansion and a reference for the reported concepts, is available on

[11] Glossary: https://www.clarin.eu/glossary?page=0

the CLARIN website. Moreover, two resources that outline subdomains of the CLARIN domain were developed within the Social Sciences & Humanities Open Cloud (SSHOC[12]) project and represented in SKOS: The SSHOC Multilingual Metadata[13] (Frontini et al., 2021b) and the SSHOC Multilingual Data Stewardship[14] (Frontini et al., 2021a) terminologies. SSHOC is an initiative funded by the EU framework programme Horizon 2020 which brings together over 20 partner organisations with the goal of advancing the social sciences and humanities sector within the European Open Science Cloud (EOSC). The project spanned over a 40-month period, from January 2019 to April 2022 with the aim to revolutionise the existing landscape of social sciences and humanities data, shifting from fragmented disciplinary divisions to a cohesive, cloud-based network that incorporates interconnected data infrastructures.

The SSHOC project was structured into 9 work packages[15], each of which delineated various tasks, deliverables, and milestones. More specifically, CLARIN was appointed as the lead beneficiary of the Work Package 3 – Lifting Technologies in the SSH Cloud. As stated on the SSHOC website, WP3 aimed to develop, enhance, and integrate tools and services essential for managing social sciences and humanities (SSH) research data, aligning with community needs and ensuring interoperability with existing functionalities. It also focused on the adaptation and enrichment of existing tools, establishing connections with the EOSC-hub e-infrastructure to facilitate sharing, and emphasising usability improvements across the SSH domain. Additionally, WP3 strived to enhance FAIRness by enabling better discovery, accessibility, and interoperability of resources. Special consideration was given to cross-disciplinary usage, such as providing language technology for SSH scenarios and survey creation and analytics technology for language resource acquisition projects.

SSHOC Multilingual Metadata (henceforth MM) and SSHOC Multilingual Data Stewardship (henceforth MDS) are the result of the SSHOC Task 3.1 "Multilingual Terminologies," aiming to explore NLP and MT approaches for the creation and fostering of multilingual SSH content across languages, improving discovery for non-native speakers. The building processes of both resources are detailed in Frontini et al. (2021).

3.6.1. SSHOC Multilingual Metadata

The SSHOC Multilingual Metadata resource is built on the metadata set of the CLARIN Concept

---

[12] SSHOC: https://sshopencloud.eu/project

[13] SSHOC Mulilingual Metadata: https://vocabs.sshopencloud.eu/vocabularies/sshocmm

[14] SSHOC Multilingual Data Stewardship: https://vocabs.sshopencloud.eu/vocabularies/sshocterm/

[15] Working Packages are listed and described here: https://sshopencloud.eu/project

Registry (CCR)[16], which forms the basis of the semantic interoperability layer of CLARIN, especially in the context of metadata, by offering a collection of concepts identifiable by their persistent identifiers. This entails that the MM Terminology is not a corpus-based resource, as the concepts derived from an already existing set of 232 concepts, which were later assigned a definition. Considering that the goal behind the implementation of this resource was to assess whether MT tools can be a successful approach in handling translation tasks, the concepts and their definitions labelled and written in English have been automatically translated into other languages, namely Dutch, French, Greek and Italian, using different MT tools. The translations have subsequently been checked and approved by native speaker experts of the relevant domain(s), in order to ensure accuracy and to determine which MT system performed better. After an analysis of the accuracy score for each translation in each language, Deep-L resulted to be the best performing MT tool, outperforming both Google Translate and Reverso (Frontini et al., 2021).

The scores also revealed that generally definitions received higher accuracy scores compared to terms, and this outcome could be explained in two ways. Firstly, translating a term within a broader context, such as a definition, is often easier because the context provides additional elements that aids the system in understanding the meaning. However, it is also worth noting that terms often have very precise and technical meanings, while definitions typically use less-specific language, making them easier to translate (Frontini et al., 2021).

3.6.2. SSHOC Multilingual Data Stewardship

The SSHOC Multilingual Data Stewardship terminology, derived from technical documentation on standards and interoperability, falls within the domain of Data Stewardship and counts a total of 211 concepts, with labels and definitions available in seven languages.

[16] CLARIN Concept Registry: https://www.clarin.eu/content/clarin-concept-registry

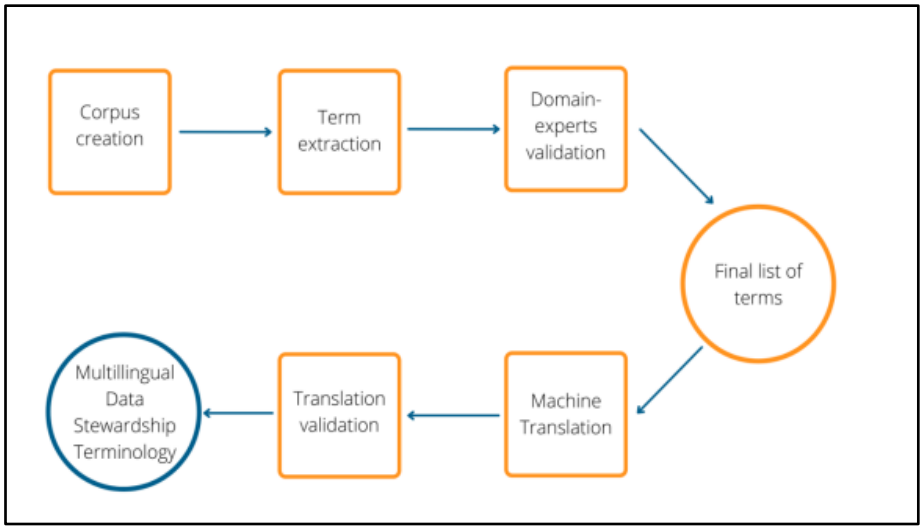

Figure 1. Workflow for the creation of the Data Stewardship Multilingual Terminology (Frontini et al., 2021)

Figure 1 shows the methodology used to create this resource, as described in Frontini et al. (2021). This methodology revolved around the usage of language technologies for the development of domain-specific terminologies, to evaluate their effectiveness, and to identify which tools produce the best results. The first step involved building a corpus, collecting 70 documents of standards and recommendations for data stewardship and curation, deliverables and other technical documents. Once the corpus was built, candidate terms were automatically extracted using four different NLP-based tools, namely SketchEngine Keywords function (Kilgariff et al., 2014), TerMine (Frantzi et al., 2000), TermoStat (Drouin, 2003) and TBXTools (Oliver and Vàzquez, 2015). After an evaluation of all the tools, the best performing proved to be TBXTools and TermoStat.

Once all the candidate terms were extracted, counting a total of 277 terms, the list went through a validation process carried out by domain experts, to assure their pertinence and relevance to the domain. The next step was the automatic translation of the concept labels (terms), and their definition into the languages of the WP partners working at the project, namely Dutch, French, German, Greek, Italian, Slovenian. It was decided to employ DeepL[17], as it turned out to be, at the time of testing, the best performing tool in the evaluation carried out for the MM Terminology. Considering that the two domains are strictly related, a good performance was expected for the domain of data stewardship as well (Frontini et al, 2021). The translation also underwent a validation process by domain experts proficient or native in the several languages included in the terminology.

For both resources, the employment of MT and NLP approaches proved to be effective in building multilingual terminologies, although the validation of the outputs by domain experts is

[17] DeepL: https://www.deepl.com/translator

fundamental to ensure reliability.

Another notable resource created within the 3.1 SSHOC project is the SKOSifying mapping tool, designed as a parser to convert the MM and the MDS terminologies. This tool transformed the flat table structure of the multilingual terminologies into SKOS format using predefined mapping rules (Trupiano, Concordia, 2021).

3.7. Comparison of SKOS resources for SSH and Open Science

In the process of determining the optimal structure to adopt for the CLARIN Thesaurus, a thorough analysis and comparison of existing SKOS vocabularies was considered essential, due to the multitude of variables that can influence the classification structure. In fact, despite SKOS offering a predefined set of properties and best practices guidelines, the organisational aspects of vocabularies can differ significantly among one another. This variation is evident in the arrangement of concepts, the employment of properties, and is often influenced by the vocabulary's intended purpose and the scale of the domain it encompasses. The following subsections examines three SKOS resources with domains closely related to or aligned with the subdomains of the CLARIN Thesaurus, to gain valuable insights for the development of a robust and effective structure.

3.7.1. Loterre Open Science Thesaurus

The first KOS under consideration is the Loterre Open Science Thesaurus[18], which delineates the primary concepts of Open Science. Developed and maintained by the *Institut de l'information scientifique et technique* (INIST), part of France's CNRS centre, this thesaurus is integrated into Loterre (Linked Open Terminology Resources),[19] a platform dedicated to the sharing of multidisciplinary terminological scientific resources (Loterre Home Page). Loterre adheres to linked open data standards and upholds the FAIR principles, namely Findability, Accessibility, Interoperability, and Reusability[20]. As part of this framework, the Open Science Thesaurus is indeed multilingual, with concepts and definitions labelled in French, English and Spanish.

This resource holds particular relevance for the CLARIN infrastructure, and by extension, for the implementation of the CLARIN Thesaurus. Like many other research infrastructures, CLARIN is committed to supporting Open Science practices and adhering to the FAIR principles.

[18] Loterre Open Science Thesaurus: https://skosmos.loterre.fr/TSO/fr/?clang=en

[19] Loterre: https://www.loterre.fr/presentation/

[20] GO FAIR: https://www.go-fair.org/fair-principles/

Consequently, the domain of Open Science stands as one of the subdomains within the broader scope of CLARIN. Furthermore, leveraging a foundational resource such as the Loterre Thesaurus aligns with the goals of the Semantic Web.

The starting points for this resource were the terms and concepts belonging to a small pool of existing glossaries, specifically wikidata.org, w3id.org, rdfs.org, semanticscience.org, and purl.obolibrary.org. The concepts in the Loterre Open Science Thesaurus are mapped to the corresponding concepts defined in these glossaries through the `skos:exactMatch` property. Additionally, the structure of the thesaurus is inspired by the taxonomy proposed by the FOSTER[21] project, which undertakes the dissemination of knowledge about Open Science and aids to develop strategies from implementing Open Science practices. Subsequently, this resource was further enriched through a text-mining process, to retrieve reference documents within the realm of Open Science.

The thesaurus comprises a total of 493 concepts systematically organised under a polyhierarchical structure consisting of 16 top concepts (Figure 3.), each branching out into complex structures. As explained in the previous chapter, in the presence of a polyhierarchy, SKOS allows concepts not only to have multiple child concepts like a simple hierarchy, but to have multiple parent concepts as well. For instance, the concept *open data* is both a narrower concept under the top concepts *data type* and *open science*, and a broader concept for linked *open data*, *open big data*, *open citation*, *open government data*, *open research data*, and *open source code*. As visible in Figure 4., all hierarchical relationships are displayed in the concept's tab.

[21] FOSTER taxonomy: https://www.fosteropenscience.eu/about#download

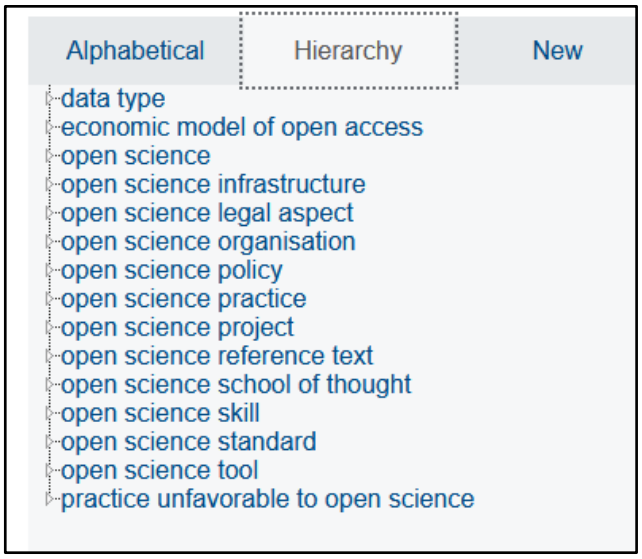

Figure 3. Top concepts of Loterre Open Science Thesaurus

Regarding the other semantic relationships, the property `skos:related` is used to establish associative relations, and associative relations are present through the record of all synonymous labels for each concept.

In terms of documentation properties, only the `skos:definition` property is employed, to report a definition for each concept and, where available, its source.

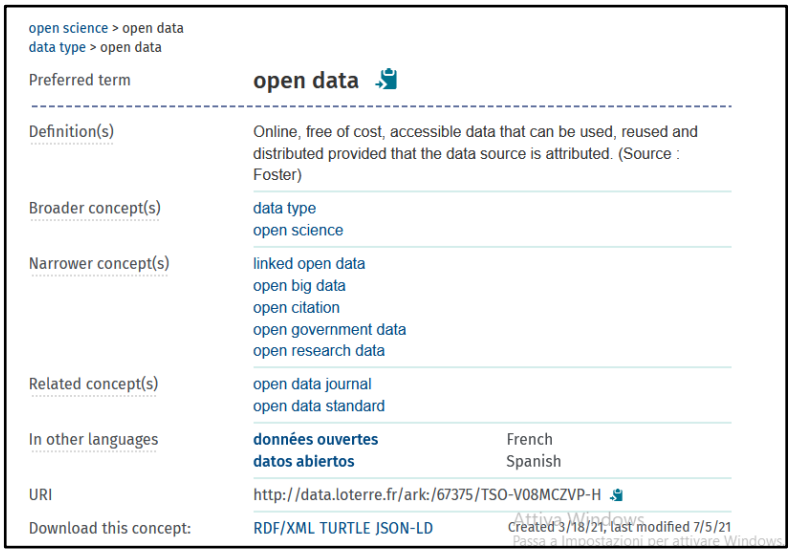

Figure 4. Example of concept in Loterre Open Science Thesaurus

3.7.2. TaDiRAH

The second terminological resource taken into consideration is TaDiRAH the Taxonomy of Digital

Research Activities in the Humanities[22], developed in a joint effort of DARIAH[23], the European Digital Research Infrastructure for the Arts and Humanities, and DiRT[24], the digital humanities tool directory. Like CLARIN, DARIAH is an ESFRI infrastructure within the field of Social Cultural Innovation. Moreover, the two infrastructures collaborate in several projects, such as SSHOC. For this reason TaDiRAH represents a valuable resource for CLARIN.

The primary objective of this knowledge organisation system is to enhance academic credibility and increase visibility for researchers in the digital humanities by offering structured terminology and knowledge regarding research activities and associated methods (Borek et al., 2016). To ensure the taxonomy's reusability for other projects, the digital humanities community was actively engaged in the resource's structuring through rounds of public feedback, providing valuable insights and suggestions. This participatory approach led to significant alterations and improvements in the taxonomy and has also been adopted for further developments, leading to significant alterations and improvements in the taxonomy. For this reason, it represents a point of reference for the development of the CLARIN Thesaurus, which could benefit from the involvement of the CLARIN community, whether for its future expansion into other languages or for any other contributions, collaborations, and opportunities for reuse.

The initial version of the taxonomy[25], created in 2014, was originally provided only in English and was divided into three sets: “Research Activities”, “Research Techniques”, and “Research Objects”. The “Research Activities” set presents a hierarchical structure consisting of eight top-level categories, each containing narrower concepts. In contrast, “Research Techniques” and “Research Objects” merely group and list their instances alphabetically, without any semantic relationships established among them.

The current version[26], released in 2020, has expanded and now includes 168 concepts and their translations in French, German, Portuguese, Serbian, Spanish, and Italian, although only at the concept level. It has also been formalised in SKOS and features a revised structure and semantics, focusing primarily on research activities and techniques. The objective was to convert the list of research activities, research techniques, and research objects utilised in the digital humanities field up to that point into an interoperable resource accessible in a machine-readable format. This resource

---

[22] TaDiRAH: https://tadirah.info/

[23] Digital Research Infrastructure for the Arts and Humanities: https://www.dariah.eu/

[24]Digital Research Tools: https://digitalresearchtools.pbworks.com/w/page/17801672/FrontPage

[25]TaDiRAH Version 1.0: https://vocabularyserver.com/tadirah/en/index.php

[26]TaDiRAH Version 2.0: https://vocabs.dariah.eu/tadirah/en/

would enable the search and retrieval of pertinent information by semantically linking the data (Borek et al., 2021). In addition, a structured representation of the data and their relationships is fundamental to showcase the applicability and range of the developed model.

In comparison to the previous version, the set of concepts referred to as "Research Objects" has been omitted, due to the lack of reliable systematic data available for modelling purposes. All concepts are identified as part of a concept scheme, and similarly to the Loterre Open Science Thesaurus, they are arranged using a polyhierarchical structure. There are seven top concepts (Figure 5.) that correspond to the research activities outlined in the previous classification, except for 'meta-activities', which are no longer included in the current version. Beneath each research activity, various research methods and techniques are listed.

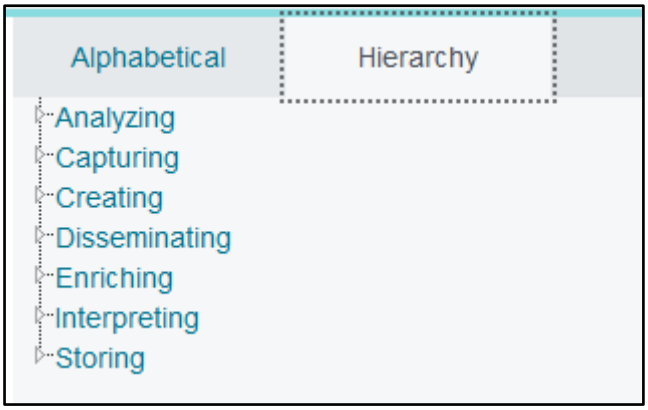

Figure 5. Top concepts of TaDiRAH

As reported in (Borek et al., 2021), the significance of the concepts is expressed through the use of two notation properties, namely `skos:scopeNote` and `skos:definition`. The former is applied when the information about the concept is written by the authors of the resource, and typically pertains to concepts that have a specialised meaning within the domain of digital humanities, despite their broader meaning in other contexts. On the other hand, the latter is utilised when providing a link to the corresponding Wikidata[9] definitions. The only other type of mapping present is expressed through the property `skos:closeMatch`, which links the current version with the first version of TaDiRAH.

Regarding the types of relationships used to describe the domain, only hierarchical relations are present, with associative and equivalent relations being entirely absent. Specifically for the latter, each concept includes only the preferred label, with no synonyms or variants listed.

3.7.3. DHA Taxonomy

The next terminological resource that was analysed is the DHA taxonomy[27], a knowledge organisation system designed to describe subjects, resources, and tools within activities of Digital Humanities Austria[28], the virtual network for the dissemination of the digital paradigm in Austrian humanities studies. It is the result of a joint effort of the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) and the Universities of Vienna and Graz. Given that ACDH-CH leads the Austrian Consortium, the DHA taxonomy holds particular relevance for CLARIN. Moreover, these collaborative endeavours are deeply intertwined with Austria's implementation of the ESFRI roadmap, as well as its commitment to European infrastructures like CLARIN-ERIC and DARIAH.EU.

The taxonomy is monolingual and counts 114 concepts each accompanied by a definition specified using the `skos:definition` property, with any available sources listed separately. Regarding the semantic relationships, associative relations are not present, as opposed to equivalence and hierarchy. Equivalence is indicated, when necessary, through the use of alternate labels. Meanwhile, the hierarchical structure is relatively linear, with most concepts situated at the same level. Some concepts have narrower concepts, maximum at a second level.

This taxonomy is taken into account especially because it offers a dual organisation of the domain. In fact, beyond the hierarchical structure, instances are also categorised into six groups, using the SKOS collection class (Figure 6.). Both collection labels and their instances are listed alphabetically.

Figure 6. DHA Taxonomy collections

Each collection has a dedicated tab where various properties are utilised, including the preferred label and scope note to delineate the content of the collection. As discussed in the previous chapter, the collection labels are not treated as individual concepts, and the URI assigned to them represents the

[27] DHA Taxonomy: https://vocabs.dariah.eu/dha_taxonomy/en/

[28] DHA: https://digital-humanities.at/en

entire collection. Consequently, all labels representing the six collections, do not exist as concepts in the hierarchical organisation. Additionally, due to the aggregation of concepts into collections, each concept is assigned the property `skos:memberOf`.

Another difference with Loterre Open Science Thesaurus and TaDiRAH is the employment of the notation property, which assigns to each concept a number that uniquely identifies it within a concept scheme. For broader concepts the notation code is a whole number, while for narrower concepts is a decimal. For example, the notation number of the broader concept ‘NLP’ is 48, while the one assigned to ‘NER’, one of its six narrower concepts, is 48.05.

4. Development of the CLARIN Thesaurus

4.1. Chapter overview

This chapter centres on the creation of the CLARIN Thesaurus. Firstly, it details the initial version of the resource developed during the internship at CLARIN ERIC, providing a comprehensive report of the methodology employed and the resulting outcomes. Successively, the focus shifts towards the systematic organisation, into a structured thesaurus, of the concepts included in the first version.
From the outset, the SSHOC Multilingual Data Stewardship terminology (Frontini et al., 2021a) has served as a significant guiding resource for the development of the CLARIN Thesaurus. Given the effectiveness of the methodology implemented for its creation, which combines automated tools with the knowledge and expertise of domain experts, it was deemed a suitable approach for the development of the CLARIN Thesaurus. Especially considering that Data Stewardship could be regarded as one of the several subdomains that make up the infrastructure, it is reasonable to assume that the employed tools would likely return satisfactory outcomes for the CLARIN domain as well.

4.2. Version 1.0.

As mentioned above, the CLARIN Thesaurus was first developed as an internship project, with the aim to obtain an English-Italian terminological resource that could serve as a base for a multilingual Thesaurus related to the CLARIN infrastructure and the services and tools it provides.

The approach adopted for the development of CLARIN Thesaurus aligns with the one employed for the SSHOC Multilingual Data Stewardship, hence the use of automation tools for term extraction and translation in other languages, both phases followed by a round of validation by domain experts. The following sections will provide a detailed description of each step of the process.

4.2.1. Corpus creation

The initial step involved the creation of a corpus containing texts relevant to CLARIN. The macro-domain of the corpus is the entire CLARIN ERIC infrastructure. However, given that CLARIN is a digital infrastructure for language-based research, the corpus encompasses various sub-domains, including, but not limited to, data management, open science practices, and language resources and technologies.

The corpus was compiled using Sketch Engine, sourcing links from both the CLARIN ERIC and the K-centres' websites. The process involved utilising the ‘find texts on the web’ feature in Sketch Engine, with URLs provided as input. URLs were obtained through the content search feature on the CLARIN website, using names of CLARIN services (e.g., Virtual Language Observatory,

Resource Families, and Language Resources Switchboard) and keywords from the list of K-Centre description tags[29] (e.g. digitisation, data mining, and discourse analysis). The output of these searches were hyperlinks to connected pages on the website. Approximately 30-50 URLs were saved for each keyword using the Linkclump web extension, resulting in a total of 550 URLs. Sketch Engine automatically extracted the texts and compiled them into a single corpus file containing 256,843 tokens.

The resulting corpus is diverse, including various text types such as research papers, informational pages, and K-Centres' pages, all presented in English. However, given the extensive nature of the CLARIN network, it is crucial to acknowledge that the corpus is not exhaustive, as it does not encompass all CLARIN-related material. Therefore, there is potential for future expansion of the corpus.

4.2.2. Extraction of candidate terms

The next phase involved the curation of a list of candidate terms, comprising both single and multi-word expressions which, as mentioned above, was carried out predominantly automatically using the corpus' keyword list, generated by Sketch Engine through an automatic comparison of the corpus with a reference corpus, according to the chosen parameters. As shown in Figure 7., the focus is on less frequent terms, to ensure the retrieval of specialised terms. The minimum frequency is set to ten, to leave out function words and hapaxes, i.e. words that occur only one time in a corpus (Faloppa, Treccani). Additionally, some terms were identified by examining the concordances of individual words within the keyword list. For example, examining the concordances of the word 'data' revealed several candidate terms, such as 'data collection,' 'data lifecycle,' and 'data curation.' A similar pattern emerged with the word 'text,' leading to the extraction of terms like 'text mining' and 'text normalisation'. The result of this process was a list of 133 candidate terms.

[29] List of K-Centre description tags: https://vonweber.nl/cgi/kcentres_atags.cgi?all

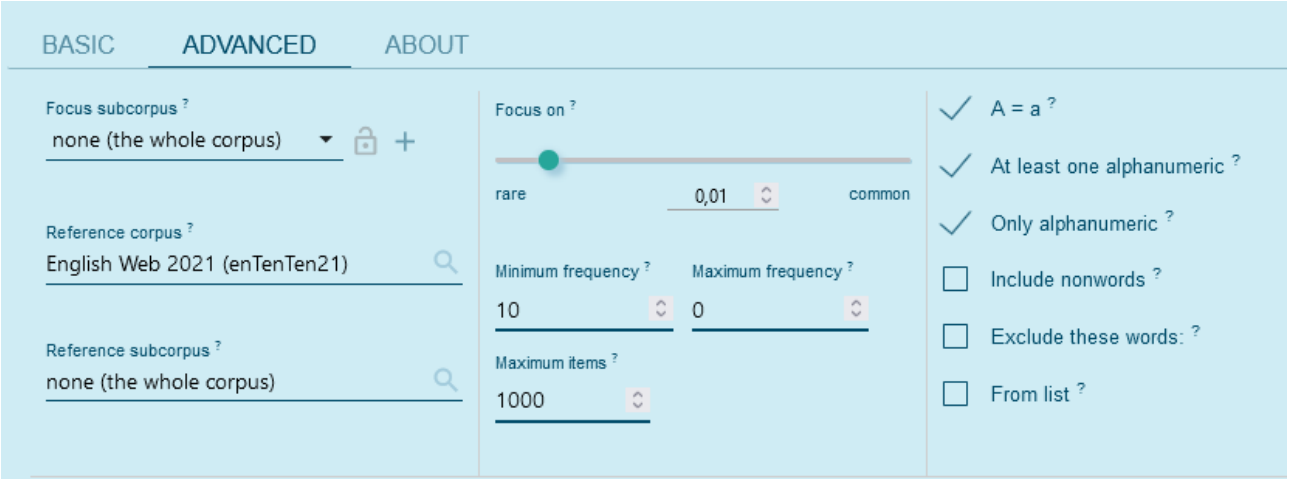

Figure 7. Keywords extraction parameters on SketchEngine

4.2.3. Validation of candidate terms

To validate the extracted candidate terms, two preliminary phases were required to prepare the document for the initial round of validation by domain experts. The validation process is essential for assessing the accuracy of the candidate terms.

The first step consisted in checking the presence of each candidate term in three resources: SSHOC Multilingual Metadata (Frontini et al., 2021b), SSHOC Multilingual Data Stewardship (Frontini et al., 2021a), and list of K-Centre description tags. In the case of the first two resources, the goal was to identify the terms that were already included in other terminologies, with the perspective to link them if the intended meaning aligned. If a term was present, the form was noted down, to keep track of whether it matched exactly or if spelling variations or synonyms were preferred. The definitions were also registered to potentially reuse them in the definition writing phase. A higher number of matching terms were found in the MDS glossary than in the MM resource, which was unexpected, since the latter has a closer relation to CLARIN, as it was based on the metadata set of the CLARIN Concept Registry. The list of K-centres tags was taken into consideration as a mean to validate the relevance of the extracted terms and to check the spelling variations.

The following step was to find an example in context for each candidate term from the created corpus. This was done using the GDEX[30] function on Sketch Engine, which stands for "Good Dictionary Examples". This function evaluates all the occurrences of a given term in a corpus in respect of their suitability to be employed as dictionary examples or for teaching purposes. It ranks them according to their length, use of complicated vocabulary, sufficient context, and presence of controversial topics (Figure 8.).

[30]GDEX: https://www.sketchengine.eu/guide/gdex/

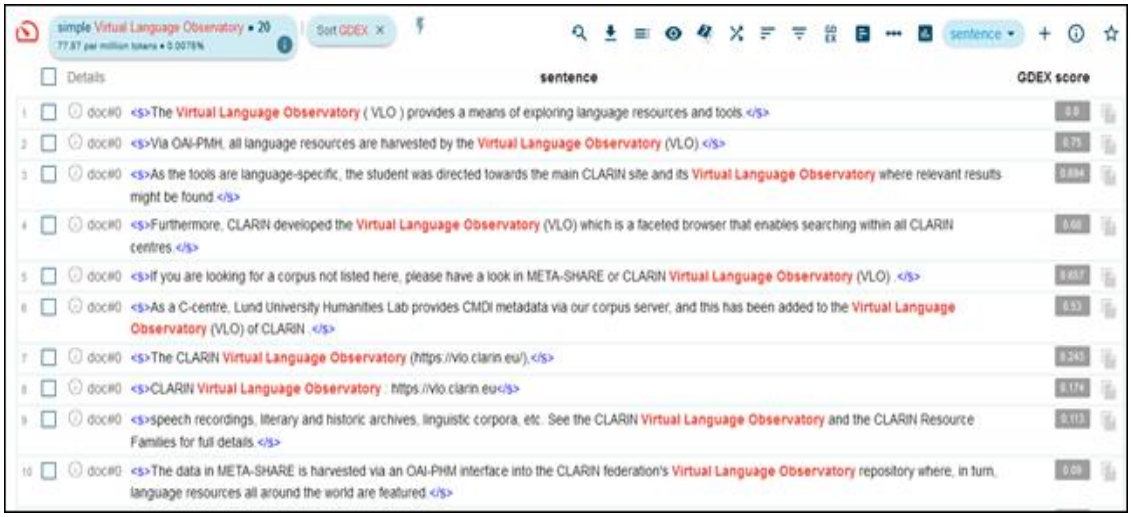

Figure 8. Example of Sketch Engine's GDEX function

The next phase consisted in the validation of the candidate terms by domain experts. Three validators were selected among members of the CLARIN community. Each validator received two files: the validation file, containing the list of candidate terms along with their corresponding examples in context retrieved in the previous phase, information collected during the comparison phase, and the validation column; the guidelines file containing information about the project's objective and methodology up to that point, and instructions on how to perform the validation.

Their task was to answer the question "Is the term, as used in the given example, part of CLARIN's domain, which includes research and linguistic data infrastructure, as well as its services and tools?" The possible answers were yes, no, and maybe. Each validator was asked to assess the relevance of the candidate terms individually to prevent any mutual influence. Once all three validators had performed the evaluation, their responses were compared to determine which terms to exclude and to assess the level of agreement.

A score was assigned to each possible answer: Yes = 1; maybe = 0.5; no = 0. It was determined that every term with a score equal or higher than 2,5 would be automatically included, those with a score lower or equal to 1 would be automatically excluded, and those with a score between 1 and 2,5 would be examined on a case-by-case basis.

This decision was influenced by the evaluation of one validator who adopted different criteria from the others:

- Yes, for terms specific to CLARIN and directly related to the infrastructure, such as services and the government bodies;
- Maybe, for terms that falls within the sub-domains of CLARIN, for example those related to data management and language technologies;

- No, for names of specific centres, projects or products;

These criteria led to the decision to include a 'label' column in the resource. This column is used to sort the concepts according to the sub-domain they pertain to: ‘CLARIN core’, ‘Open Science’, ‘LRT’. This aspect will be further explored in the upcoming sections (See Sections 3.2.7 and Chapter 4.), as it represents the basis for the expansion of the thesaurus.

To establish the level of agreement or reliability between the validators, Cohen’s Kappa was calculated. Cohen’s Kappa (*k*) is a statistical measure to assess the inter-rater agreement when analysing categorical or nominal data. It allows us to determine whether agreement between the raters is more than what would be expected by chance alone. As one of the three validators adopted different criteria for the task, it was decided to perform the analysis considering only two validators, as it was clear that including the third one would have resulted in a lower score not necessarily reflecting the reliability of the other two validators. To calculate Cohen’s Kappa the data need to be organised in a contingency table, showing the frequency of agreement and disagreement between the raters. Table 3. reports the frequency of each combination of assessment. The diagonal of the table represents the instances where the validators agree, while the off-diagonal entries represent cases where the validators disagree. In this case, it can easily be noted that the level of agreement is significantly high, with 122 instances out of 133, all obtained with the category ‘yes’.

	Validator 2			
Validator 1	Yes	Maybe	No	Total
Yes	122	2	2	126
Maybe	6	0	1	7
No	0	0	0	0
Total	128	2	3	133

Table 3. Validators assessment

The formula for the calculation of Cohen’s kappa is:

$$k = (p_o - p_e) / (1 - p_e)$$

Where p_o is the relative observed agreement among raters and p_e is the hypothetical probability of chance agreement. The observed agreement alone is not enough to establish the true level of agreement, because it includes instances where the validators agreed by chance. Therefore it is

necessary to calculate the agreement by chance (p_e) and take it out of the equation.
A *k* value of 1 indicates perfect agreement between raters, while a *k* value equal or minor to 0 indicates no agreement. With the data provided in Table 3., the *k* value for the validation of the extracted terms resulted to be 0.57, which indicates moderate agreement between the two validators.

Eventually, after the validation only a few candidate terms were excluded, while all the remaining were later reorganised according to the three identified groups. Overall, the validation by domain experts has been extremely valuable as each validator provided precious insights and advice, for example the addition of certain terms for completion's purposes.

4.2.4. Establishment of semantic relations

In the field of Knowledge Organization Systems (KOSs), three primary types of relationships are recognized (See Section 2.7.1.): equivalence, association, and hierarchy. While all three were considered for this resource, it was ultimately decided that a flat representation would suffice for the time being, as the primary objective was to develop a bilingual terminological resource, with less emphasis on the systematic organisation of the CLARIN domain. Consequently, concepts are displayed alphabetically rather than hierarchically. Nevertheless, some relationships were identified to clarify and disambiguate certain concepts.

Regarding equivalence, synonyms and alternate forms were identified. Corpus searches were conducted to determine the preferred label for each concept and, if present, alternate labels, based on their frequency. For example, the term selected as the preferred label to denote "a type of CLARIN centre that specialises in language resources and provides knowledge and expertise to researchers and educators from all linguistic fields" is 'K-Centre', with 'Knowledge Centre' serving as the alternate label. Acronyms, typically used as alternate labels, are predominantly adopted as preferred labels in this resource, reflecting their common usage by users.
Another aspect worth noting regarding preferred and alternate labels is the distinction between UK and US spelling. When a concept is found in the corpus with both UK and US variants, the UK spelling is consistently set as the preferred label. This decision is based on the standard usage of British spelling within CLARIN, and it is recommended that all users adhere to this convention. Therefore, certain CLARIN entities, such as 'CLARIN Centre', will always be labelled with UK spelling ('centre') rather than US spelling ('center'), even if the latter is used in some publications, as the former is considered the technical term.

Regarding hierarchy, as mentioned in Section 3.3.2, SKOS allows for the establishment of generic hierarchical relationships without specifying the type, thereby giving users the flexibility to determine the nature of these hierarchical relationships. To maintain simplicity, only genus/species relationships (A is a type of B) were considered, as this is the most common form of hierarchy.

As for association, this relation was used to indicate that a given concept or instance is closely related to another, but not in a hierarchical way. For instance, the concepts *CMC* (Computer-Mediated Communication) and *CKCMC* (CLARIN Knowledge Centre for CMC and Social Media Corpora) are identified as related: while CKCMC is a K-Centre specialising in Computer-Mediated Communication, their relationship is non-hierarchical.

4.2.5. Writing of definitions

After organising the concepts semantically, the next step was to formulate definitions. These definitions were taken from reliable sources and adapted when necessary.

Given CLARIN's commitment to adhering to the FAIR principles, particularly emphasising reusability, the primary sources chosen for drafting definitions were pre-existing terminological or lexical resources. The primary sources included IATE, the multilingual terminology database developed by the European Union; the CLARIN ERIC website, especially for concepts belonging to the CLARIN core category; and the Oxford Reference[31] website, which offers a diverse range of accessible specialised Oxford dictionaries, such as *A Dictionary of Computer Science (7 ed.)*[32] (Butterfiled et al., 2016) or *A Dictionary of the Internet (4 ed.)*[33] (Ince, 2019). Given the fundamental importance of the FAIR principles and Open Science for the operational effectiveness of the infrastructure, it was considered prudent to primarily utilise existing terminologies and resources. This approach facilitates mapping and aligning with established standards, thus ensuring adherence to the principle of reusability.

In instances where a suitable definition was not available in any of these resources, alternative sources such as academic papers or other field-specific websites, such as NLPlanet[34] and Towards Data Science[35] were consulted.

This phase resulted to be the most laborious and time-consuming, as determining the correct definition was often challenging. In fact, information from different sources sometimes contradicted each other, and it happened to encounter almost identical definitions for different concepts. Moreover, significant effort was dedicated to this phase, as the definitions constitute a crucial component of the

---

[31] Oxford Reference: https://www.oxfordreference.com/

[32] A Dictionary of Computer Science: https://www.oxfordreference.com/display/10.1093/acref/9780199688975.001.0001/acref-9780199688975

[33] A Dictionary of the Internet: https://www.oxfordreference.com/display/10.1093/acref/9780191884276.001.0001/acref-9780191884276

[34] NLPlanet: https://www.nlplanet.org/

[35] Towards Data Science: https://towardsdatascience.com/

resource, especially considering their role as source text for translation into other languages.

4.2.6. Translation

As mentioned in the introduction, the translation of the concepts and their definitions into Italian was carried out automatically, with the chosen MT system being DeepL. The output was successively post-edited, and overall this approach significantly expedited the translation process.

Automatic translation returned satisfactory results for the definitions, which required only a moderate amount of post-editing, primarily at the syntax level, as the automatic translation tended to be more literal and maintain the Italian syntax, which does not sound natural in English. One tendency was that the longer the definition, the less accurate the translation, and the greater the amount of revision needed. On the other hand, closer attention was necessary for the translation of terms. Approximately 25% of the terms required some form of post-editing. In many cases, the automatic translation was not incorrect, but a different variant was preferred in Italian (Examples 1 and 2). This occurred especially with acronyms, where the output coincided with the input, and while the same acronym is used in Italian, the expansion is preferred (Examples 3 and 4). Another common issue arose with concepts that, in Italian, are referred to by their English term. This was particularly prevalent with named entities (Examples 5 and 6), as well as instances where literal translations resulted in incorrect terms (Examples 7 and 8). To ensure that the terms in Italian were accurate, it was crucial that a native Italian CLARIN expert validated the Italian translations of the terms.

	English term	**Automatic translation**	**Post-Edited translation**
Example 1	virtual collection	collezione virtuale	raccolta virtuale
Example 2	C-Centre	Centro C	Centro CLARIN di tipo C
Example 3	HLT	HLT	Human Language Technology
Example 4	VLO	VLO	Virtual Language Observatory
Example 5	CLARIN Newsflash	Notizie CLARIN	CLARIN Newsflash
Example 6	CLARIN workshop	Laboratorio CLARIN	Workshop CLARIN

Example 7	deep learning	apprendimento profondo	deep learning
Example 8	dependency parsing	analisi delle dipendenze	dependency parsing

Table 4. Examples of post-edited terms

4.2.7. Conversion into SKOS format

Once the spreadsheet containing the resource was finalised, it underwent conversion into SKOS format using the conversion tool specifically developed for the MDS terminology. (Trupiano, Concordia 2021). This tool employs YAML[36] data serialisation language to map the columns in the spreadsheet with SKOS properties.

SKOS has been chosen as the appointed representation format for various reasons. First of all, SKOS is endorsed as the recommended format in task 3.5 "Data and Metadata Interoperability" within D3.1 "Report on SSHOC (meta)data interoperability problems" (Broeder et al., 2019), primarily due to its capability to assign a URI to each concept, thereby enhancing interoperability and reusability. Moreover, SKOS facilitates the organisation of concepts in a straightforward and vertical manner, which was deemed suitable for the objectives of the present study. Another reason lies in its flexibility: while there are guidelines outlining best practices, these are not rigid restrictions, allowing users the freedom to determine the structure and properties (See Table 2.) to utilise based on the resource's purpose and the domain requirements.

As regards to the structure of the resource, as briefly mentioned in Section 4.2.4., it was determined that, at least for the time being, a flat representation would suffice, meaning that the concepts are consultable only in alphabetical order and the overall domain is not presented in a structured and well-defined hierarchy. This entails that concepts have not been systematically organised into a concept scheme or grouped into collections. Hierarchical relations of the genus/species type have been established between concepts, however, to maintain a flat representation, only broader concepts are displayed for narrower ones, using the `skos:broader` property , and not vice versa. For instance, the concept *K-Centre* is classified as a narrower term under *CLARIN Centre* because it denotes a specific type of CLARIN Centre. This hierarchical relationship is presented in the *K-Centre* entry, whereas the broader concept *CLARIN Centre* does not display its narrower concept in its own tab.

[36]YAML: https://yaml.org/

Considering label properties, preferred and alternate labels have been employed to express equivalence, while the chosen documentation properties are `skos:definition`, `skos:note`, `skos:editorialNote`, `skos:scopeNote`:

- `skos:definition` was used for the English definition retrieved from reliable sources and its automatic translation into Italian. The source of the definition is displayed using the `dc:source` property;
- `skos:note` was used for any additional information about the concept, written in English and then automatically translated into Italian;
- `skos:editorialNote` was used to signal the provenance, i.e. whether the concept was extracted from the corpus or if it was added afterwards for completion purposes. For instance, the term 'K-Centre' was extracted from the corpus, whereas the other types of CLARIN Centre, namely B-Centre and C-Centre, were not extracted but included afterwards. The addition was necessary to ensure the presence of all relevant related items;
- `skos:scopeNote` is usually used for information about the intended meaning of a concept, however, for this terminological resource it was employed to signal the label of the term, used to sort the concepts according to the domain they pertain to. There are three labels. The first one is 'CLARIN core', for all concepts strictly related to CLARIN and its infrastructure, such as services, governance bodies and projects or initiatives that are either funded by or related to CLARIN. The second is 'Open Science' for concepts related to FAIR data management and research. The third one is 'LRT' for general concepts within the domain of language resources and technologies. These three groups will serve as the basis for the implementation of the domain classification.

As for mapping properties, it was decided to employ only the `skos:exactMatch` property, to link concepts to other resources where the concept is included with the same form and the same meaning. The terminological resources where at least one exact match was found are: SSHOC Multilingual Metadata, SSHOC Multilingual Data Stewardship, and IATE.

Since the CLARIN Thesaurus is bilingual, for all the properties that needed string literals in both languages – namely preferred and alternate labels, definitions, and notes – a language tag consisting of 2-letter ISO language codes has been added.

Figure 9. below represents one of the concepts included in the thesaurus:

Figure 9. Example of concept in CLARIN Thesaurus

4.2.8. Results and limitations

After the conversion, the thesaurus, comprising 152 concepts, was published on a test instance of Skosmos[37]. Skosmos[38] is an open-source web-based SKOS browser and repository, that allows for the publication, visualisation and retrieval of linked data. Among these concepts, 32 were designated as belonging to the category CLARIN core, 40 to Open Science, and 78 to LRT. In two instances, a single label was deemed insufficient to categorise the concept: *SSHOC*, which is considered part of both the 'CLARIN core' and 'Open Science' domains, and *language data*, labelled as 'Open Science' and 'LRT'. These initial groupings lay the groundwork for further classification of all concepts.

The resource effectively offers an overview of the key components of the CLARIN infrastructure and the underlying technologies supporting its resources and tools. While it can already be utilised for certain applications, such as translating and localising CLARIN content into Italian, it serves merely as a foundation for others. For instance, for the topic annotation of the content available on the website, a more structured thesaurus may prove more beneficial for users, as well as for adequately representing the complex and multifaceted infrastructure of CLARIN. Moreover, to cater to a broader user base, and make the resource useful for the majority of the CLARIN community, expansion into other languages is fundamental. With regards to the included concepts, the resource lacks certain key elements that would enrich its comprehensiveness and better represent all aspects relevant to the domain of CLARIN. For instance, the term 'interoperability', representing one of the

[37] First version of the CLARIN Thesaurus: https://v4e-dock.isti.cnr.it/skosmos/pltclt/en/page/?uri=https%3A%2F%2Fclarin.eu%2Fxxx%2Fct_108_POS_tagging

[38] Skosmos: https://skosmos.org/

FAIR principles, and labelled as ‘Open Science’, was extracted from the corpus, while the remaining three principles (findability, accessibility, and reusability) were not. Similarly, within the CLARIN core group, ‘national coordinator’ was included, yet other governmental bodies are absent. Having adopted a bottom-up approach based on the data included in the corpus, this outcome was anticipated and highlights one of the drawbacks of this approach. Given that both the corpus creation and term extraction involved some degree of automation, there was a higher potential for errors or inaccuracies.

4.3. Version 2.0

While expanding the terminology into other languages is a separate process, the endeavour to incorporate additional concepts is linked with the proposal of a structured thesaurus. This approach involved organising the existing 152 concepts included in the thesaurus into a hierarchical structure and addressing the gaps with missing concepts. The three groups identified while implementing the first draft of the resource served as a starting point for the classification process.

It is important to emphasise that the primary objective of the current work is to propose a classification of the terminology related to CLARIN to enhance content retrieval on the website. The results will be handed to the governing bodies of CLARIN, who will discuss the accuracy and applicability of the CLARIN Thesaurus. The focus was the establishment of a coherent systematic organisation of concepts, starting from those extracted from the corpus. Since a bottom-up approach was adopted, the categories may initially lack some instances, as only concepts extracted from the corpus will populate them. These elements can be supplemented later by incorporating existing appropriate resources or through further expansion of the thesaurus. However, using a top-down approach, several concepts have been added, particularly for the top classes or higher nodes of the classification. In fact, the majority of these concepts were not present in the first version of the resource, but they were added to facilitate the grouping of certain concepts under a category.

For this classification, polyhierarchy is permitted, recognizing that some concepts naturally belong to multiple classes. This allows for a more nuanced and flexible organisation, accommodating the multifaceted nature of certain concepts. Moreover, a polyhierarchical structure is beneficial for content search, as it facilitates access and retrieval of information across different contexts. Some specific instances of concepts with multiple parent classes are highlighted in the subsequent sections. However, a comprehensive overview of all such cases is provided in Table D of the Appendix.

Furthermore, it is important to specify that the focus was primarily on the hierarchical organisation of the concepts, hence mainly hierarchical relationships are discussed. While associative relations were also considered during the process, they are not thoroughly addressed in the current work. Nevertheless, Table E in the Appendix aims to report all identified associative relations, which

will be implemented in the conversion into SKOS.

The following sections provide a detailed description of the organisation of concepts according to the three identified groups: CLARIN Core, LRT, Open Science. For each group, the main subclasses are discussed, addressing the encountered issues. The structured classification is represented in Protégé[39], an open-source ontology editor. Although Protégé supports SKOS as well as OWL, it was used in this work exclusively for the hierarchical representation of concepts.

4.3.1. 'CLARIN core' concepts

As specified above, this resource aims to comprehensively cover all the aspects of the infrastructure, leading to the inclusion of several knowledge sub-domains relevant to CLARIN. Some of these, such as the domain of data management, are also pertinent to other infrastructures or companies. The sole domain uniquely associated with CLARIN is the category previously labelled 'CLARIN core', which grouped concepts related to the organisation of the infrastructure, and the main services it provides, thus closer attention is placed on it.

Out of the three groups, it resulted to be the one with the smallest number of concepts, counting only 32. As mentioned in Section 3.2.8., this outcome is expected when adopting a bottom-up approach, as a crucial role is played by the corpus, which may not have been fully representative of the domain. Upon reviewing the existing concepts labelled as 'CLARIN core', alphabetically listed in Table A (Appendix), it becomes apparent that this category is rather incomplete. Therefore, a top-down approach was adopted for the addition of new concepts. The initial step involved identifying the primary classes into which the concepts could be categorised.

The identified subclasses are: *CLARIN Core Services*, *CLARIN Governance*, *CLARIN initiatives and Events*, *CLARIN Jointly Maintained Services*, and *CLARIN Structure* (Figure 10.). Table A also reports, for each concept, the subclass it belongs to. Some of the concepts are flagged as instances. Instances represent real objects pertaining to one of the categories. Handling these instances poses a significant challenge for this work, so they will be separately addressed in Section 4.4.

[39] Protégé: https://protege.stanford.edu/

Figure 10. Subclasses of *CLARIN*

The detected classes facilitated an initial hierarchical organisation of the existing concepts, with *CLARIN* positioned as the top concept. This means that while the other classes are regarded as second-level classes under the *CLARIN* concept, they serve as the top nodes for the branching structure of existing concepts. This approach allowed us to understand which concepts should have been added to the resource to enhance the representation of CLARIN as an infrastructure. None of the identified classes were previously included in the resource, so they were incorporated and defined as new concepts.

Table 5. lists all subclasses of *CLARIN* alphabetically. For each of them a definition was drafted, based on the information available on the website, to facilitate better classification.

Subclasses of *CLARIN*	**Definitions**
CLARIN Core Services	**Proposed definition:** *Services dedicated to enhancing the discoverability, interoperability, and reusability of language resources, managed primarily or entirely by the CLARIN Central Hub.*
CLARIN Governance	**Proposed definition:** *The governing bodies that manage the infrastructure, each with distinct functions but working cooperatively.*
CLARIN Initiatives and events	**Proposed definition:** *Range of Initiatives and events promoted by CLARIN aimed at fostering community engagement and enhancing user experience across various portals. This encompasses all elements listed under the 'Learn&Exchange' section on the website.*
CLARIN Jointly Maintained Services	**Proposed definition:** *Services or initiatives funded and managed by CLARIN in collaboration with other European infrastructures or institutions.*
CLARIN Structure	**Proposed definition:** *Organization of CLARIN as a distributed infrastructure*

Table 5. Subclasses of *CLARIN* and their definitions

For the placement of concepts under the identified subclasses, the information available on the CLARIN website and its segmented sections served as a crucial reference point. Given that one of the main applications of the resource is the enhancement of the website's user-friendliness and facilitation of topic retrieval, an effort was made to preserve the website's categorization as close as possible, which lead to the inclusion of several concepts that were not extracted from the corpus, but represent fundamental elements of the infrastructure. Some of the node concepts necessitated further deliberation to determine the appropriate classification and identify potential concepts for inclusion.

The first subclass is *CLARIN Core Services* (Figure 11.), which encompasses all the services available for users, directly provided by CLARIN. Some of them were not extracted from the corpus, namely *Federated Content Search*, *CLARIN Federated Login*, and *Depositing Services*, but were added for completion purposes. Generally speaking, within this subclass all the discovery services are grouped, i.e. all the services that facilitate the retrieval of content and resources within the website and infrastructure. The concept initially labelled as 'content search' in the first version has been updated to 'Federated Content Search' to more accurately reflect its relevance to the CLARIN domain.

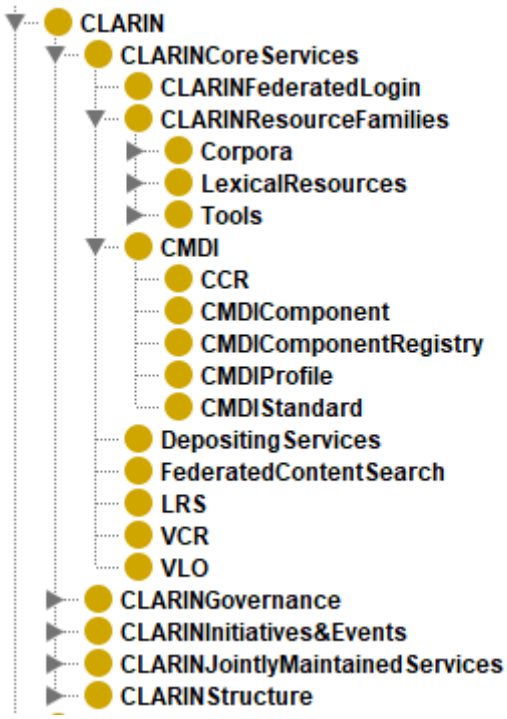

Figure 11. *CLARIN Core Services* subclasses

The second subclass to be addressed is *CLARIN Governance* (Figure 12.). Only four concepts

pertaining to this class were already included in the first version of the resource: *Board of Directors*, *National Coordinator*, *CLIC*, and *ERIC*. To add relevant concepts and establish a classification, the Governance page[40] on the CLARIN website served as the point of reference. The dedicated page explains the bodies governing the infrastructure, their functions, and their relationships, which are depicted in Figure 13. below. It was evident that several types of relationships are simultaneously in place, while the current work aimed to maintain a simple structure. Therefore, instead of establishing hierarchical relationships among the concepts, they were left at the same level, except for the five thematic committees. Associative relationships were established instead (Table E, Appendix), and the exact nature of the connections among the concepts are explained in the definitions.

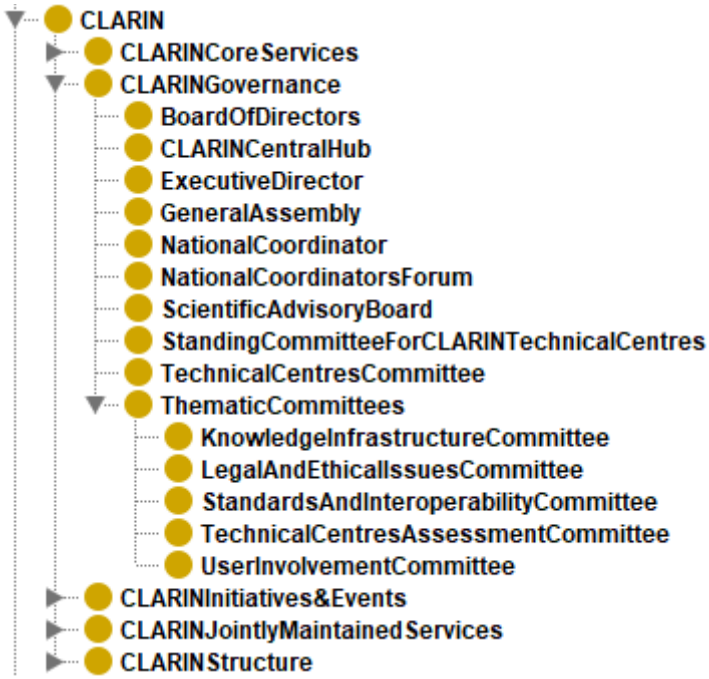

Figure 12. *CLARIN Governance* subclasses

[40] CLARIN Governance page: https://www.clarin.eu/content/governance

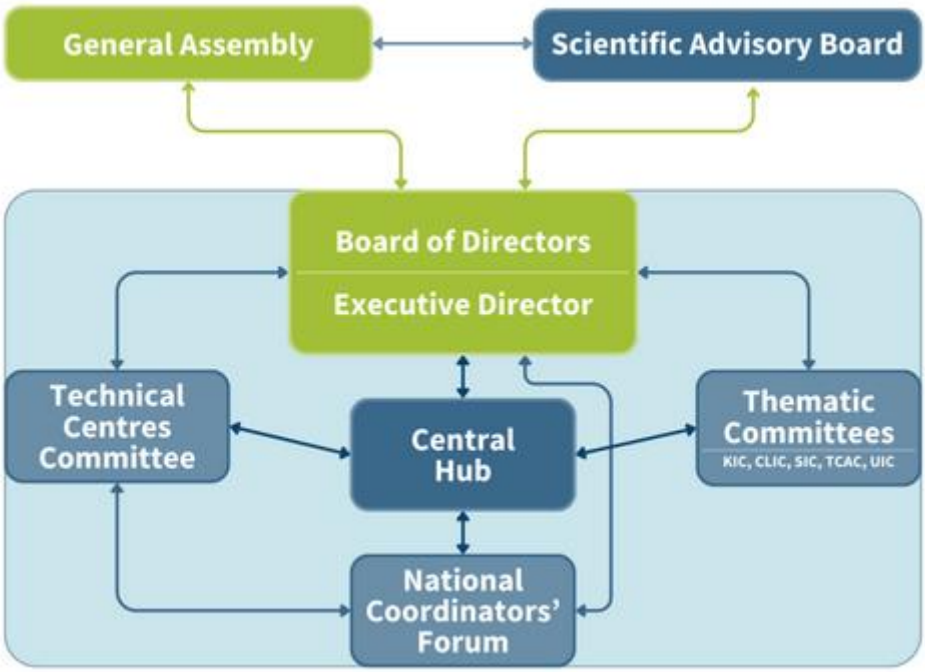

Figure 13. Visual Representation of CLARIN Governance

Among the four concepts initially classified under governance, the concept *ERIC* has been relocated, as it would have been inappropriate to categorise it as a narrower concept of CLARIN. ERIC stands for 'European Research Infrastructure Consortium' and in the first version of the thesaurus was defined as "a specific legal form that facilitates the establishment and operation of Research Infrastructures with European interest." This implies that while it may be perceived as a governing body, this concept is the one that signifies CLARIN as a European infrastructure. As a result, it was decided to add another top concept, at the same level of *CLARIN*, labelled as *European RI Landscape* (Figure 14). This class aims at providing an overview of the panorama of European Research Infrastructure. As for many other classes of this classification, the instances are not comprehensively reported, as the focus was more on the establishment of classes useful for the organisation of the concepts extracted for the first version of the thesaurus.

One of the subclasses is *EOSC*, a concept present in the first version of the thesaurus, defined as "European cloud service that hosts and processes research data and knowledge within the scientific community." A subclass of *EOSC* is *Science Cluster*, as the cloud service provides scientific data relevant to all scientific sectors, including the field of Social Sciences and Humanities (SSH). Consequently, *SSHOC* is categorised as a narrower concept of *SSH*.

Another subclass is *ESFRI*, which represents all large-scale research infrastructures of pan-European interest, covering various scientific areas. As mentioned in Section 3.5., CLARIN and DARIAH are both ESFRI infrastructures, so they are categorised as instances within this subclass.

The final subclass is *Legal Status*, which currently includes only two instances: *AISBL*, identifying an institution or entity as a non-profit organisation, and *ERIC*, specifying the European nature of a research infrastructure.

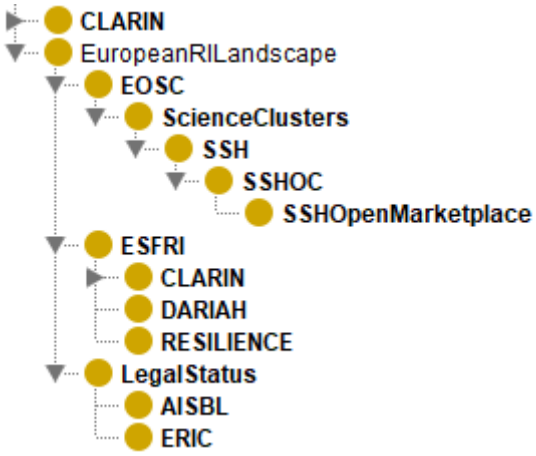

Figure 14. *European RI Landscape* subclasses

The third subclass is *CLARIN Initiatives and Events* (Figure 15.), which encompasses initiatives and other services not linked to discovery. These initiatives are divided into *CLARIN User Involvement* and *Learn & Exchange*. The *CLARIN User Involvement* category includes initiatives and events that directly engage users, such as the annual CLARIN Conference and CLARIN Workshops, which focus on topics related to improving its services or expanding its network. The CLARIN Newsflash, a newsletter that keeps users updated with the latest news regarding the infrastructure, is included as well. The *Learn & Exchange* category comprises instances grouped under this label on the website. Their common ground is that they provide access to materials or repositories for finding training materials. For instance, the Learning Hub offers access to a great variety of open educational resources on a wide range of subjects. These resources include comprehensive online training modules for skill development and materials for creating university courses, conducting training sessions, and organising workshops.

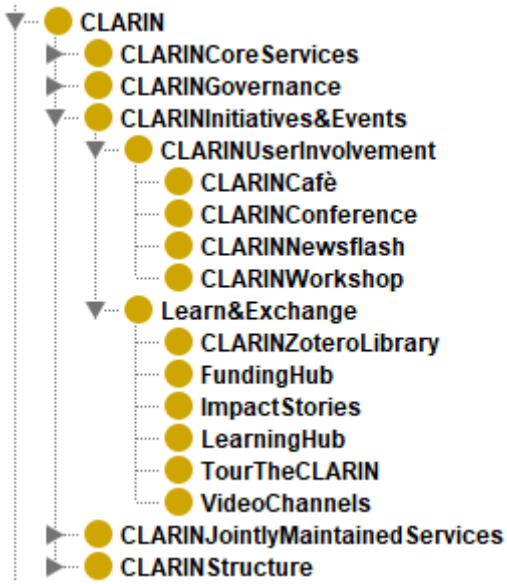

Figure 15. *CLARIN Initiatives & Events* subclasses

The fourth subclass is *CLARIN Jointly Maintained Services* (Figure 16.), which, as defined in Table 5., refers to initiatives or services that CLARIN funds and manages in partnership with other European infrastructures or organisations. The three concepts placed as part of this class are actually instances, as they represent projects or platforms within the realm of CLARIN resources. The concept *SSH Open Marketplace* was initially labelled as 'Open Science' in the first version of the thesaurus. However, considering that it is one of the key services developed within the SSHOC project, it was more appropriate to categorise it as a narrower concept of *SSHOC*.

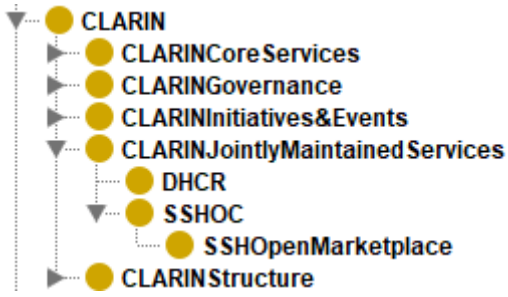

Figure 16. *CLARIN Jointly Maintained Services* subclasses

The final subclass is *CLARIN Structure* (Figure 17.), which represents the configuration of CLARIN as a distributed infrastructure. It reports the presence of a National Consortium for each member country and details the various types of centres within the infrastructure.

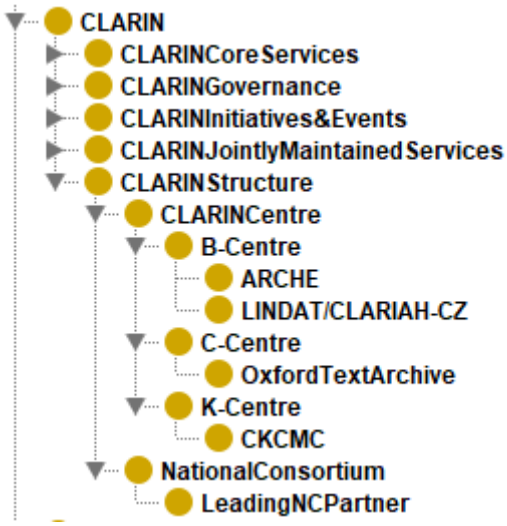

Figure 17. *CLARIN Structure* subclasses

4.3.2. 'Open Science' concepts

The second group that underwent the classification process was labelled 'Open Science' in the first version of the thesaurus, and included a total of 40 concepts related to the management of research data and the practices of Open Science (Table B., Appendix). However, upon further analysis, it was deemed inappropriate to designate 'Open Science' as the top concept, as not all concepts align strictly with its principles. For instance, while Open Science promotes the free and accessible dissemination of research data, not all aspects of research data management revolves around Open Science principles. In light of this, an effort was made to identify a more encompassing label that could accommodate all concepts. Therefore, *Research Data Ecosystem* was appointed as the top concept. This term can be defined as "the comprehensive environment that encompasses the subdomain of data management, the foundational principles underlying research activities, and the associated infrastructures." This definition aims to encompass all other identified classes, which are: *Research Data Management*, *Open Science*, *Research Infrastructure*, *Type of Data* (Figure 18.). Each of them will be delineated in the following paragraphs. Table 6. below provides the definitions for each subclass. Among them, only *Type of Data* was not part of the first version of the thesaurus.

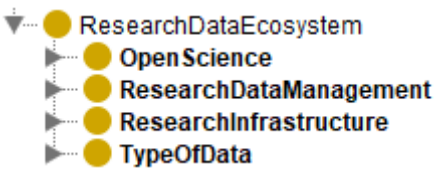

Figure 18. *Research Data Ecosystem* subclasses

Subclasses of *Research Data Ecosystem*	Definitions
Research Data Management	the disciplines related to managing research data as a valuable resource.
Open Science	movement to make scientific research, data and dissemination accessible to all the members of a research community.
Research Infrastructure	facility that provides resources and services for the research communities to conduct research and foster innovation in their fields.
Type of Data	**Proposed definition**: *various forms of data based on their characteristics*

Table 6. Subclasses of *Research Data Ecosystem* and their definitions

The first subclass is *Open Science* (Figure 19.), for which the Loterre Open Science Thesaurus serves as the primary reference, as most concepts falling under this category are also featured in the Loterre Open Science Thesaurus, which offers a more comprehensive coverage of this domain. For this reason, it was deemed appropriate to mirror the structure and labels of the Loterre Thesaurus as closely as possible to ensure reusability. For instance, all the subclasses of *Open Science*, except for *Open Access* and *Open Data*, which are concepts already present in the thesaurus, are acquired from the sixteen top concepts of the Loterre thesaurus. As specified in the introduction of the chapter, the classes may appear incomplete, as only the concepts extracted for the first version are included. The *Open Science* class could be integrated and mapped with the Loterre Open Science Thesaurus in a subsequent version of the thesaurus.

Another aspect of the Loterre thesaurus adopted for the CLARIN thesaurus is its organisation under a polyhierarchical structure, wherein concepts can have multiple parent concepts. This decision is particularly suitable for the domain of CLARIN, as many concepts are interconnected, and establishing more hierarchical structures can enhance content retrieval. For example, *FAIR Data* is classified as a narrower concept under both *Open Science Guidelines* (Figure 19.) and *Type of Data* (Figure 22.). This means that the branching of concepts starting from *FAIR Data* will be displayed under both parent classes.

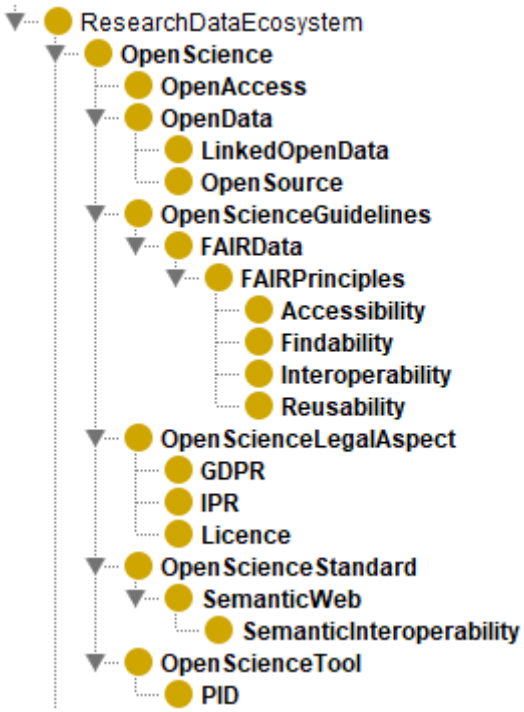

Figure 19. *Open Science* subclasses

The second subclass is *Research Data Management* (Figure 20.). This concept could be considered narrower than *Data Management*, which is not included in this domain, because not necessarily the practices of data management revolve around research data. Narrower of this concept is *Data Infrastructure*, which is defined in the first version of the CLARIN thesaurus as "the underlying technological and organisational components that support the collection, storage, processing, and usage of data." This means that this is a broader concept that includes all the concepts that refer to the processing of data, specifically covering all steps of the data lifecycle, as outlined in the proposed structure.

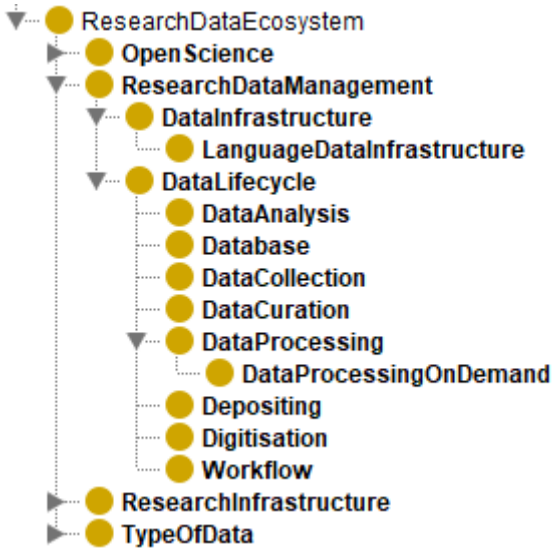

Figure 20. *Research Data Management* subclasses

The third subclass is *Research Infrastructure* (Figure 21.), which, as defined in Table 6., refers to the legal entity or facility that provides data for the research community. *Data Infrastructure* is also considered a narrower concept within this category, resulting in having two parent concepts. Other narrower concepts under *Research Infrastructure* are *research community*, which encompasses all individuals working and interacting with the infrastructure, and *research data repository*, which refers to databases provided by the infrastructure to ensure data accessibility.

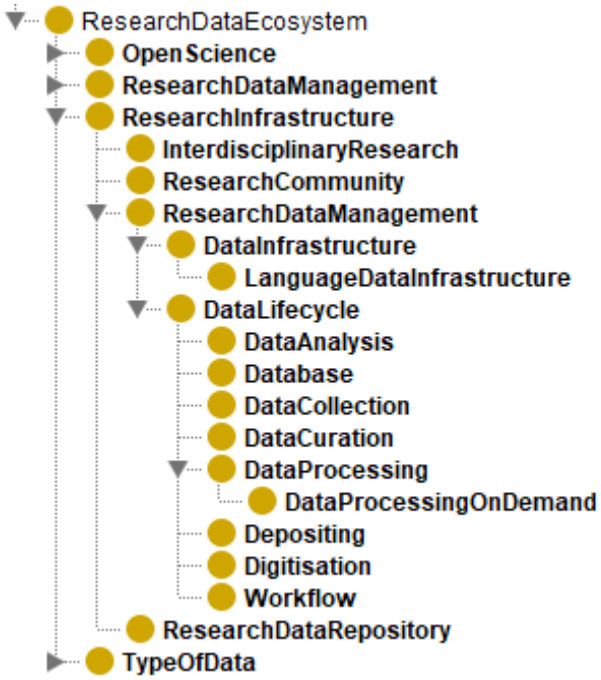

Figure 21. *Research Infrastructure* subclasses

The final subclass is *Type of Data* (Figure 22.), which is adapted from the Loterre Open Science Thesaurus, where the label used is 'Data Type'. The designation 'Type of Data' was preferred for its broader applicability, considering the diverse concepts included in this class. For example, *language data* and *textual data* are quite general, making a wide-ranging label more suitable.

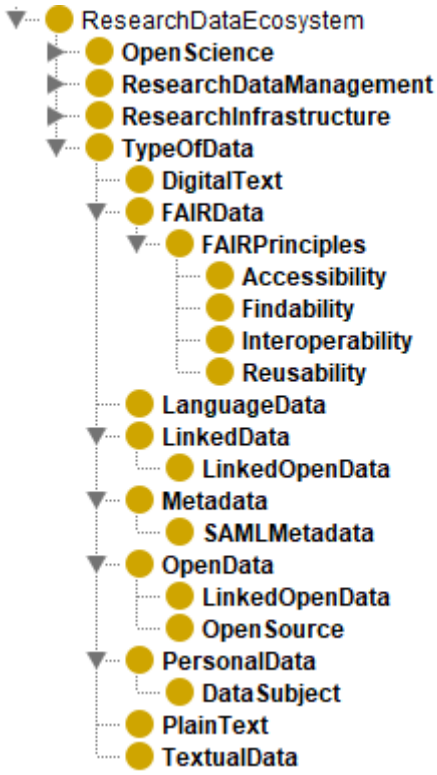

Figure 22. *Type of Data* subclasses

4.3.3. 'LRT' Concepts

The third category is the LRT group, which includes concepts related to language resources as well as the tools and techniques used for processing and analysing language data across different research fields. In the initial version of the resource, 79 concepts were sorted in the LRT group, detailed in Table C (Appendix). Initially, 'LRT' was used only as a grouping term for these concepts and did not represent the preferred label for any specific concept. Subsequently, it was decided to define and adopt the expanded acronym 'Language Resources and Technologies' as the preferred label. This term was included as a top concept, defined as "a comprehensive domain encompassing computational tools, techniques, and language resources that enable linguistic analysis in support of multidisciplinary research."

For the hierarchical organisation of these concepts, an important point of reference was the DHA Taxonomy which describes the panorama of language resources and activities for Digital Humanities Austria. As reported in Section 3.5.3., this taxonomy presents a dual organisation of the concepts: hierarchical and categorical. For the latter, six groups are established to gather all the concepts, defined through the `skos:collection` class. As recommended in the SKOS reference (Miles, Bechhofer, 2009), these groups are not considered concepts. These categories (Figure 6., Section 3.5.3.) inspired the selection of the subclasses of the LRT group, which include *Disciplines*, *Practitioners*, *Resources*, *Standards*, *Techniques*, and *Tools* (Figure 23.). Despite the DHA taxonomy

not treating collection labels as concepts, it was decided for the current project to define them as concepts to maintain consistency throughout the thesaurus with the treatment of the top nodes. Table 7. reports on the subclasses of *Language Resources and Technologies* and their proposed definitions.

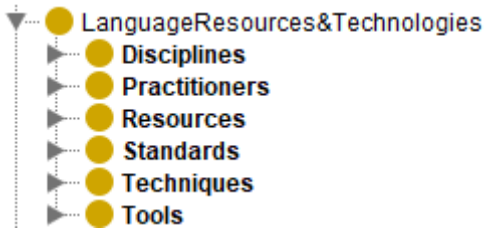

Figure 23. *Language Resources and Technologies* subclasses

Subclasses of *Language Resources and Technologies*	**Definitions**
Disciplines	**Proposed definition**: *All knowledge fields relevant to the CLARIN infrastructure.*
Practitioners	**Proposed definition**: *All professionals working within the fields supported by CLARIN.*
Resources	**Proposed definition**: *All language resources developed or utilised within CLARIN for language-based research.*
Standards	**Proposed definition:** *All standards supported and endorsed within the CLARIN infrastructure*
Techniques	**Proposed definition**: *All methods used for language analysis and processing within CLARIN.*
Tools	**Proposed definition**: *All instruments used for language analysis and processing within CLARIN.*

Table 7. Subclasses of *Language Resources and Technologies* and their definitions

The first subclass is *Disciplines* (Figure 24.), which groups all subject fields involved in CLARIN. Most of the concepts are set at the same level. Exceptions are the multidisciplinary fields *AI*, *computational linguistics*, *corpus linguistics*, which encompass further subcategories. The concepts *natural language processing pipeline* and *collocation* have been placed under *NLP* and *corpus linguistics* respectively. Although they do not directly belong to the *Disciplines* category, their hierarchical relationship to the broader concepts is justifiable for retrieval purposes.

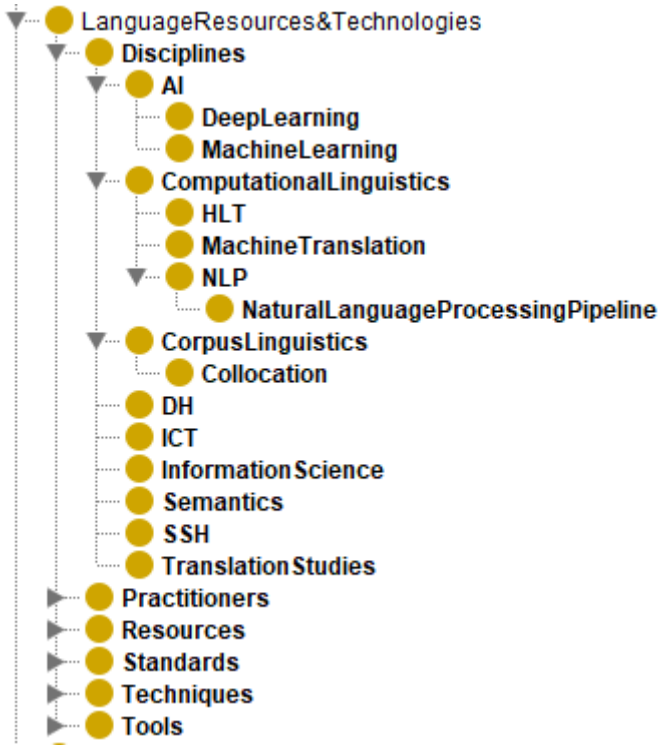

Figure 24. *Disciplines* subclasses

The second subclass is *Practitioners* (Figure 25.), consisting solely of two elements: *linguist* and *computational linguist*. Placing these concepts posed a challenge because it would be conceptually inaccurate to classify them as narrower concepts under *computational linguistics* or *corpus linguistics*, as their relationship is associative rather than hierarchical. Also, in this context, they represent rather general concepts, especially *linguist*, but an effort was made to include all concepts extracted and validated from the corpus.

Also, considering that the main application for the CLARIN thesaurus is the annotation of content on the website, general terms like 'linguist' are not particularly relevant, but it was kept anyway as it could be appropriate for other purposes and applications. For example, CLARIN is currently working on the normalisation of the vocabulary used in reporting. Reporting is crucial in a large digital infrastructure like CLARIN, as it helps keep the community informed about achievements, regulations, and other important updates. Various working groups and committees are responsible for reporting on different aspects of the infrastructure. However, each committee tends to use its preferred terms, leading to a wide variety of synonyms and variants, which causes ambiguity. A set of terms that exhibits significant variation pertains to the classification of disciplines and practitioners. For this reason, it was considered appropriate to create a dedicated class for practitioners, even though it currently only includes two narrow concepts. This class could be further populated and its instances, together with concepts pertaining to the *Disciplines* class, could then be

used to create a comprehensive list of discipline-practitioner pairs. It is important to remark that current work serves as a proposal for the systematic organisation of concepts related to the CLARIN infrastructure and will be discussed internally with the governing bodies.

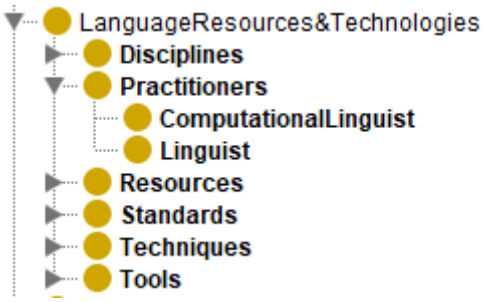

Figure 25. *Practitioners* subclasses

The third subclass is *Resources* (Figure 26.), which encompasses various types of language and text resources. This subclass is further divided into four additional categories: *corpora*, *language resources*, *training materials*, and *virtual collections*. As is later reported in Section 4.4, the concepts *corpora* and *lexical resources* have two parent concepts, namely *Resources* and *CLARIN Resources Families*. Hence, all their child concepts are displayed in both ramifications.

The fourth class, *Standards* (Figure 26.), currently includes only one concept, *TEI*. However, it can be expanded in future iterations to incorporate additional standards supported or curated by CLARIN, particularly those managed by the CLARIN Standards and Interoperability Committee (SIC).

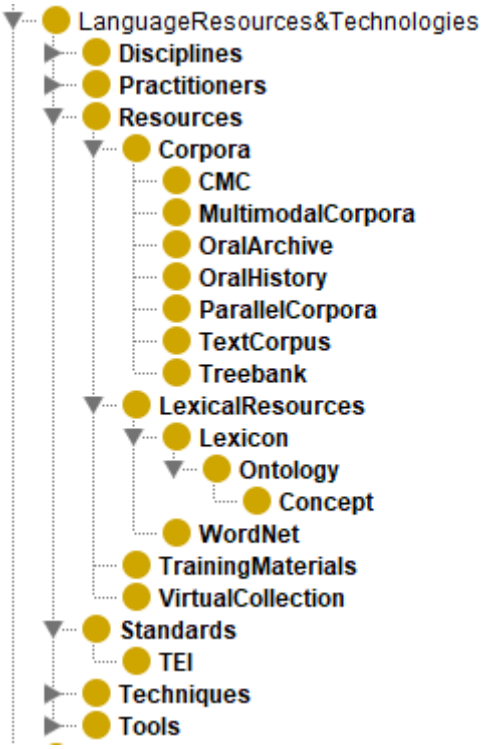

Figure 26. *Resources* and *Standards* subclasses

The fifth subclass, *Techniques* (Figure 27.), and the sixth subclass, *Tools* (Figure 28.), are treated separately and each has its own ramifications. However, many of the concepts are interconnected, with all the tools, such as annotation tools and parsers, being designed to perform specific techniques. For *techniques*, most concepts are positioned at the same level, with just a few exceptions. For instance, *annotation* is further divided into its various types such as automatic and manual annotation. Likewise, *text processing* is a broader and general concept that includes a wide range of techniques like PoS-tagging, parsing, tokenization, etc.

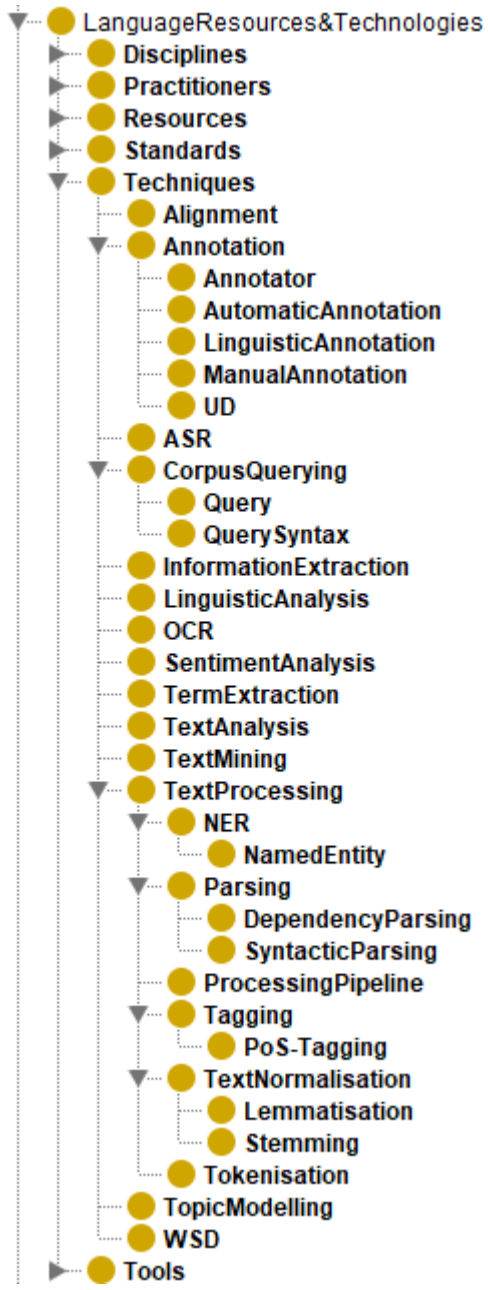

Figure 27. *Techniques* subclasses

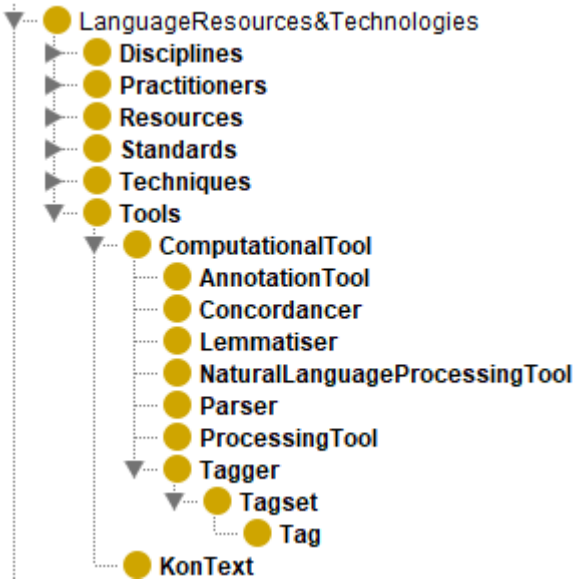

Figure 28. *Tools* subclasses

4.4. Challenges and future works

This section addresses the primary challenges encountered during the classification process of concepts present in the CLARIN Thesaurus.

As noted in Section 4.1, one of the primary and anticipated limitations of the initial version of the resources was the insufficiency of concepts extracted from the corpus for a comprehensive representation of the CLARIN infrastructure domain. This resulted in the omission of certain classes or elements, which needed to be identified and included. However, given the vast scope of the infrastructure and the wide range of subdomains, adding all missing concepts would have been a laborious task. Consequently, only concepts within the 'CLARIN core' group, directly related to the infrastructure, were systematically added. The other two groups can be integrated with existing resources, such as the Loterre Open Science Thesaurus for the 'Open Science' group. This choice was also given by the fact that the main objective of the current work was the classification of all the concepts already present in the first version of the resource into a coherent structure, rather than the addition of new concepts.

Another significant issue was the presence of instances of classes, often represented by named entities, which are real objects denoted by proper name. These instances, along with their assigned definitions in the first version of the CLARIN Thesaurus, and the class they belong to, are reported in the table below (Table 8.). The ideal approach would have been the inclusion of all the instances for all the identified categories. However, given the main objective, it was considered appropriate to include only the instances already present in the first version of the resource. This issue could be addressed in future developments by automatically populating categories with existing resources that group all the instances. For example, the CLARIN Centre Registry[41] lists all the centres in the infrastructure, detailing the consortium they belong to and the services they provide, thus identifying their type (K, B or C). This registry could be used to integrate all centres under their respective types.

[41] CLARIN Centre Registry: https://centres.clarin.eu/

Concept	Definition	Broader concept
ARCHE (A Resource Centre for the HumanitiEs)	Resource centre founded by the Austrian Centre for Digital Humanities and Cultural Heritage that provides depositing services and easy access to digital language resources	B-Centre
CKCMC (CLARIN Knowledge Centre for CMC and Social Media Corpora**)**	K-Centre that offers expertise on language resources and technologies for Computer-Mediated Communication and Social Media.	K-Centre
LINDAT/CLARIAH-CZ	Collaborative research infrastructure in the Czech Republic, formed by merging LINDAT/CLARIN and DARIAH-CZ, which focuses on language data and various digital resources, offering them to researchers, industries, and the public and provides tools, services, and technologies for language resources and digital data processing.	B-Centre
KonText	basic query interface developed by the Czech National corpus for working with corpora.	Tools
Oxford Text Archive	one of the CLARIN's centre that provides repository services for literary and linguistic datasets.	C-Centre

WordNet	lexical database of English that organises words and their meanings into a structured network of interconnected sets of synonymous words or concepts called synsets.	Lexical Resources
EOSC	European cloud service that hosts and processes research data and knowledge within the scientific community.	European RI Landscape
SSH Open Marketplace	project funded by the EU framework programme Horizon 2020 and unites 20 partner organisations and their 27 associates in developing the social sciences and humanities area of EOSC.	SSHOC

Table 8. Concepts categorised as Instances

Another challenge arose with the subcategories of *CLARIN Resource Families*, one of CLARIN's discovery services. Resource Families[42] are defined on their dedicated page as an "overview per data type of the available language resources in the CLARIN infrastructure for researchers from the digital humanities, social sciences, and human language technologies." The various types of resources are organised into three main categories, namely *Corpora*, *Lexical Resources*, and *Tools*. As shown in Figure 29., all types of resources available for use in CLARIN are listed under each category.

[42] CLARIN Resource Families page: https://www.clarin.eu/resource-families

Figure 29. CLARIN Resource Families

The issue with these three categories resides in the fact that they were already included in the first version of the terminology as belonging to the LRT group, considering their general meaning. For example, the definition of *Corpora* as a narrower concept of *CLARIN Resource Family* could be "Family of resources available in CLARIN that encompasses all different types of corpora, which are large and structured collections of written, spoken, or recorded texts." While the definition of *corpora* as a concept pertaining to LRT is simply "large and structured collection of written, spoken, or recorded texts." The same occurs with several other terms referring to types of resources, all belonging to the LRT group. For instance, the concept *multimodal corpora* is defined in the CLARIN Thesaurus as "data collections that include diverse forms of communication like text, speech, images, and more, helping to study how humans communicate through different senses and modes," representing the general meaning of the concept. However, within the context of CLARIN Resource Families, this label is used to indicate the presence of this type of resource among the data types provided by CLARIN. In other words, it refers to the application of this concept within the CLARIN infrastructure. Therefore, including *multimodal corpora* as a narrower concept of *Corpora* (narrower of *CLARIN Resource Families*) would either require altering the existing definition of the concept or having two concepts with the same label, leading to ambiguities. As the general approach was not to include all instances for each class, not all elements pertaining to the three resource families are directly included. Nevertheless, as polyhierarchy is allowed for this classification, *Corpora*, *Lexical Resources*, and *Tools* have two parent concepts: other than *CLARIN Resource Families*, the first two are also found under *Resources*, while *Tools* represent one of the main subclasses of the *Language Resources and Technologies* branch. This entails that all child concepts placed under these three concepts will be displayed in both ramifications. However, as opposed to the two other Resource

Families, with *Tools* this approach causes inconsistencies between the current classification and the official classification of Resource Families (Figure 29.). In fact, the elements that are classified as tools on the website are classified as techniques in the current work, and consequently, those classified as tools in the current work do not appear on the website.

Regarding future developments, beyond integrating instances with other existing resources, the primary focus will be expanding the thesaurus into additional languages, achieving the overarching goal of creating a multilingual resource. The approach used for the Italian translation can be replicated, involving the use of machine translation (MT) systems to translate the terms and their definitions, followed by post-editing and validation by CLARIN experts who are native speakers of the target languages.

Moreover, the thesaurus will undergo subsequent developments during a post-graduate internship at the Universidad Politécnica de Madrid in collaboration with the Ontology Engineering Group (OEG). The internship will focus on the exploration of various RDF models to convert existing terminological resources, including the CLARIN Thesaurus, into other semantic web formalisms such as Ontolex-Lemon.

5. Conclusion

The current work revolved around the creation of the CLARIN Thesaurus, a bilingual terminological resource that encompasses all concepts pertaining to the domain of the CLARIN infrastructure and the services and tools it offers. The main application of this resource is the topic annotation of content on the CLARIN website, to optimise information retrieval. The website is the access point to the infrastructure, but its distributed nature and the absence of a central systematic regulation of the specialised vocabulary lead to a considerable diversity in the terminology, which inevitably affects the effectiveness of content retrieval.

Building on the flat resource developed as an internship project, the current work focused on organising the extracted concepts into a coherent hierarchical structure based on the three identified subdomains: 'CLARIN core,' 'Open Science,' and 'LRT.' Having adopted a bottom-up approach, it became clear that the initial set of extracted concepts was insufficient for a comprehensive representation of the domain. Consequently, additional classes were identified and added to enhance the overall structure. Beyond the classes, the tendency was to not add all missing concepts, as the priority was the organisation and placement of the already extracted concepts. Moreover, the 'LRT' and 'Open Science' groups represent broader subdomains that have already been documented in resources such as the Loterre Open Science Thesaurus. Therefore, these groups could be integrated with existing resources to enhance their representation. Nevertheless, for the 'CLARIN core' group, a slightly different approach was adopted. Given its central role in the resource, particular attention was dedicated to making this group as comprehensive as possible and closely mirroring the sections of the website. This led to the inclusion of several concepts, such as all CLARIN governing bodies.

Another aspect that required careful consideration was the handling of instances, often represented by named entities. In the first version of the thesaurus, some instances were extracted, but many classes remain incomplete. Since the project's primary focus was to organise the already extracted elements and establish a coherent structure, it was decided to place only the extracted instances under their corresponding classes. The missing instances can be added in subsequent developments of the thesaurus, potentially through the integration of other resources that already group them.

The classification that resulted from the current work is merely a proposal, which has to be revised and validated by the governing bodies of CLARIN ERIC. For the moment, the updated Thesaurus, like its first version, will be converted into SKOS and published on the test instance of Skosmos.

References

Bellandi, A., Di Nunzio, G. M., Piccini, S., & Vezzani, F. (2023) The Importance of Being Interoperable: Theoretical and Practical Implications in Converting TBX to OntoLex-Lemon. *Proceedings of the 4th Conference on Language, Data, and Knowledge*, 646-651.

Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., & Summers, E. (2013) Key choices in the design of Simple Knowledge Organization Systems (SKOS). *Journal of Web Semantics, 20*, 35-49. https://doi.org/10.1016/j.websem.2013.05.001.

Berners-Lee, T., Fielding, R., & Masinter, L. (2005) Uniform Resource Identifier (URI): Generic Syntax. https://datatracker.ietf.org/doc/html/rfc3986.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001) The Semantic Web: A New Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New Possibilities. In O. Seneviratne & J. Hendler (Eds.), *Linking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web* (pp. 91-103). Association for Computing Machinery. https://doi.org/10.1145/3591366.3591376.

Borek, L., Dombrowski, Q., Perkins, J., & Schöch, C. (2016) TaDiRAH – A Case Study in Pragmatic Classification. *Digital Humanities Quarterly, 10*(1). https://www.digitalhumanities.org/dhq/vol/10/1/index.html.

Borek, L., Hastik, C., Khramova, V., Illmayer, K., & Geiger, J. (2021) Information Organization and Access in Digital Humanities: TaDiRAH Revised, Formalized and FAIR. *Proceedings of the 16th International Symposium on Information Science*, 321-332. https://dblp.org/db/conf/isiwi/isiwi2021.html.

Bowker, L., & Pearson, J. (2002) *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge.

Cabré, M. T. (1999) *Terminology: Theory, methods and applications*. John Benjamins Publishing Company.

Cimiano, P., McCrae, J. P., & Buitelaar, P. (2016) *Lexicon Model for Ontologies: Community Report*. WC3 Community Group. https://www.w3.org/2016/05/ontolex/.

Depecker, L. (2015) How to build terminology science? In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology. Volume 1* (pp. 34-44). John Benjamins Publishing Company. https://benjamins.com/catalog/hot.1.how1.

Drewer, P. & Schmitz, K. D. (2017) *Terminologiemanagement: Grundlagen - Methoden - Werkzeuge*. Berlin/Heidelberg: Springer Vieweg. https://link.springer.com/book/10.1007/978-3-662-53315-4.

Drouin, P. (2003) Term extraction using non-technical corpora as a point of leverage. *Terminology, 9*(1), 99-115. https://doi.org/10.1075/term.9.1.

Montiel-Ponsoda, E., Bosque-Gil, J., Gracia, J., Aguado de Cea, G., & Vila-Suero, D. (2015) Towards the Integration of Multilingual Terminologies: An Example of a Linked Data Prototype. In *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence*, 205-206. https://dblp.org/db/conf/tia/tia2015.html.

Fišer, D. & Witt, A. (2022) *CLARIN: The Infrastructure for Language Resources*. Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783110767377,

Frantzi, K., Ananiadou, S., & Mima, H. (2000) Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries, 3*(2), 115-130. https://link.springer.com/journal/799/volumes-and-issues/3-2.

Frontini, F., Gamba, F., Monachini, M., Broeder, D., Tijdens, K., & Vipavc Brvar, I. (2021) *D3.9 Report on Ontology and Vocabulary Collection and Publication*. Zenodo. https://doi.org/10.5281/zenodo.5913485.

Gilchrist, A. (2003) Thesauri, Taxonomies and Ontologies: An Etymological Note. *Journal of Documentation, 59*(1), 7-18. https://www.emerald.com/insight/content/doi/10.1108/00220410310457984/full/html.

Grigoris, A. & van Harmelen, F. (2004) A Semantic Web Primer. Cambridge (MA), USA: MIT Press. https://mitpress.mit.edu/9780262012102/a-semantic-web-primer/.

Gruber, T. R. (1995) Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies, 43*(5-6), 907-928. https://www.sciencedirect.com/science/article/pii/S1071581985710816.

Gruber, T. R. (1993) A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199-220. https://www.sciencedirect.com/science/article/pii/S1042814383710083.

Harping, P. (2010) *Introduction to controlled vocabularies – Terminology for Art, Architecture, and Other Cultural Work*. Los Angeles (CA), USA: Getty Publications. https://www.getty.edu/publications/virtuallibrary/160606018X.html.

Heylen, K. & De Hertog, D. (2015) Automatic Term Extraction. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology*, Volume 1 (pp. 203-221). Amsterdam, Netherlands: John Benjamins Publishing Company. https://benjamins.com/catalog/hot.1.aut1.

Nakagawa, H. & Mori, T. (2002) A Simple but Powerful Automatic Term Extraction Method. *COLING-02: COMPUTERM 2002: Second International Workshop on Computational Terminology*. https://aclanthology.org/W02-1407/.

Hjørland, B. (2007) Semantics and Knowledge Organization. *Annual Review of Information Science and Technology, 41*(1), 367-405. https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/aris.2007.1440410115.

Hodge, G. (2000) *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Alexandria (VA), USA: Council on Library and Information Resources. http://www.clir.org/pubs/reports/pub91/contents/.

Hyvönen, E. (2002) The Semantic Web – The New Internet of Meanings. In E. Hyvönen (Ed.), *Semantic Web Kick-Off in Finland - Vision, Technologies, Research, and Application* (pp. 3-20). Helsinki, Finland: Helsinki Institute for Information Technology. https://research.aalto.fi/en/publications/semantic-web-kick-off-in-finland-vision-technologies-research-and.

Isaac, A. & Summers, E. (2009) *SKOS Simple Knowledge Organization System Primer*. World Wide Web Consortium. https://www.w3.org/TR/skos-primer/.

ISO 1087:2019. *Terminology work and terminology science*.

ISO 30042:2019 *Management of terminology resource – TermBase eXchange (TBX)*.

Jacquemin, C. (1999) Syntagmatic and Paradigmatic Representations of Term Variation. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 341-348. https://aclanthology.org/P99-1044/.

Kageura, K. (2015) Terminology and lexicography. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology*, Volume 1 (pp. 45-59). Amsterdam, Netherlands: John Benjamins Publishing Company. https://benjamins.com/catalog/hot.1.ter2.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014) The Sketch Engine: ten years on. *Lexicography, 1*(1), 7-36. https://link.springer.com/article/10.1007/s40607-014-0009-9.

Klyne, G. & Carroll, J. J. (2004) *Resource Description Framework (RDF): Concepts and Abstract Syntax – W3C Recommendation*. World Wide Web Consortium. https://www.w3.org/TR/rdf10-concepts/.

Koivunen, M. R. & Miller, E. (2002) W3C Semantic Web Activity. In E. Hyvönen (Ed.), *Semantic Web Kick-Off in Finland - Vision, Technologies, Research, and Application* (pp. 27-41). Helsinki, Finland: Helsinki Institute for Information Technology. https://research.aalto.fi/en/publications/semantic-web-kick-off-in-finland-vision-technologies-research-and.

Lockinger, G., Kockaert, H. J., & Budin, G. (2015) Intensional definitions. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology*, Volume 1 (pp. 60-81). Amsterdam, Netherlands: John Benjamins Publishing Company. https://benjamins.com/catalog/hot.1.int1.

Magris, M., Musacchio, M. T., Rega, L., & Scarpa, F. (2002) (Eds.) *Manuale di terminologia. Aspetti teorici, metodologici e applicativi*. Milano: Hoepli. https://www.hoeplieditore.it/universita/articolo/manuale-di-terminologia-marella-magris/9788820329433/0157.

Matthews, B. (2005) Semantic web technologies. *E-Learning and Digital Media, 6*(6). https://www.researchgate.net/publication/30408878_Semantic_Web_Technologies.

Mazzocchi, F. (2018), Knowledge Organization System (KOS). *Knowledge Organization, 45*(1), 54-78. https://www.isko.org/ko.html.

McCrae, J.P., Gil, J.B., Gràcia, J., Bitelaar, P., & Cimiano, P. (2017) The OntoLex-Lemon Model: Development and Applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference* (pp. 587-597). https://elex.link/elex2017/proceedings-download/.

McGuinness, D. L. & van Harmelen, F. (2004) *OWL Web Ontology Language Overview – W3C Recommendation*. World Wide Web Consortium. https://www.w3.org/TR/owl-features/.

Miles, A. & Bechhofer, S. (2009) *SKOS Simple Knowledge Organization System Reference – W3C Recommendation*. World Wide Web Consortium. https://www.w3.org/TR/skos-reference/.

Miles, A., Rogers, N., & Beckett, D. (2004) *SKOS-Core 1.0 Guide. An RDF Schema for thesauri and related knowledge organisation systems.* World Wide Web Consortium. http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990) Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography, 3*(4), 235-244. https://academic.oup.com/ijl/article-abstract/3/4/235/923280?redirectedFrom=fulltext.

Nakagawa, H. & Mori, T. (1998) Nested Collocation and Compound Noun for Term Extraction. In *Proceedings of the First International Workshop on Computational Terminology* (pp. 64-71).

Oliver, A., & Vàzquez, M. (2015) TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 473-479). https://aclanthology.org/R15-1062/.

Cimiano, P., McCrae, J. P., Rodríguez-Doncel, V., Gornostay, T., Gómez-Pérez, A., Siemoneit, B., & Lagzdins, A. (2015) Linked terminologies: applying linked data principles to terminological resources. In *Electronic lexicography in the 21st century: linking lexical data in the*

digital age. Proceedings of the eLex 2015 conference (pp. 504-517). https://elex.link/elex2015/conference-proceedings/paper-34/.

Pieterse, V. and Kourie, D. G. (2014) Lists, Taxonomies, Lattices, Thesauri and Ontologies: Paving a Pathway through a Terminological Jungle. *Knowledge Organization, 41*(3), 217-229. https://www.isko.org/ko.html.

Reineke, D. (2014) TBX between termbases and ontologies. *Terminology and Knowledge Engineering 2014*. https://hal.science/hal-01005838v2.

Roturier, J. (2019), XML for translation technology. In M. O'Hagan (Ed.), *The Routledge Handbook of Translation and Technology*. Milton Park, UK: Taylor & Francis. https://doi.org/10.4324/9781315311258.

Santos, C. & Costa, R. (2015) Domain specificity: Semasiological and onomasiological knowledge representation. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology,* Volume 1 (pp. 153-179). https://benjamins.com/catalog/hot.1.dom1.

Smith, A. (2022) Simple Knowledge Organization System (SKOS). *Knowledge Organization, 49*(5), 371-384. https://www.nomos-elibrary.de/10.5771/0943-7444-2022-5-371/simple-knowledge-organization-system-skos-volume-49-2022-issue-5?page=1.

Soergel, D. (2009) *Knowledge Organization Systems: Overview*. Available at http://www.dsoergel.com/SoergelKOSOverview.pdf.

Spärck-Jones, K. (1992) Thesaurus. In S. C. Shapiro (Ed.), *Encyclopedia of Artificial Intelligence,* Volume 2 (pp. 1605-1613). New York (NY), USA: Wiley.

TEI Consortium (2023) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Available at https://tei-c.org/guidelines/p5/.

Thurmair, G. (2003) Making term extraction tools usable. *EAMT Workshop: Improving MT through other language technology tools: resources and tools for building MT*. https://aclanthology.org/2003.eamt-1.20/.

Touretzky, D. S. (1986) *The Mathematics of Inheritance Systems*. Los Altos (CA), USA: Morgan Kaufmann.

Trupiano, L. & Concordia, C. (2021), *SSHOC Data Stewardship terminology and Metadata SKOSifying mapping*, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa. Available at http://hdl.handle.net/20.500.11752/ILC-566.

Tudhope, D. & Binding, C. (2008) Faceted thesauri. *Axiomathes, 18*(2), 211-222.

Weinberg, B. H. (1998) ASIS'97: The Classification Research Workshop. *Key Words, 6*(2), 21-22.

Wittgenstein, L. (1953) *Philosophical Investigations*. Translation by E. Anscombe. New York (NY), USA: Macmillan.

Wright, S. E. & Wright, L. D. (1997) Terminology Management for Technical Translation. In S. E. Wright & L. D. Wright (Eds.), *Handbook of Terminology Management*, Volume 1. Amsterdam, Netherlands: John Benjamins Publishing Company. https://benjamins.com/catalog/z.htm1.19wri.

S. E. Wright & L. D. Wright (Eds.), *Handbook of Terminology Management*, Volume 1: *Basic Aspects of Terminology Management*. Amsterdam, Netherlands: John Benjamins Publishing Company. https://benjamins.com/catalog/z.htm1.

Wüster, E. (1931) *Internationale Sprachnormung in der technik besonders in der Elektrotechnik*. Düsseldorf, Germany: VDI Verlag.

Zeng, M. L. (2008) Knowledge Organization Systems (KOS). *Knowledge Organization, 35*(2-3), 160-182. https://www.nomos-elibrary.de/10.5771/0943-7444-2008-2-3-160/knowledge-organization-systems-kos-volume-35-2008-issue-2-3?page=1.

Language resource references

ACDH-OEAW. (2015) DHA Taxonomy. https://vocabs.acdh.oeaw.ac.at/dhataxonomy/dhaTaxonomyScheme.

Borek, L., Hastik, C., Khramova, V., Geiger, J. (2020) TaDiRAH Version 2.0: https://vocabs.dariah.eu/tadirah/en/

Digital Research Tools. https://digitalresearchtools.pbworks.com/w/page/17801672/FrontPage

Frontini, F., Gamba, F., Monachini, M., & Broeder, D. (2021a) *SSHOC Multilingual Data Stewardship Terminology*, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa. Available at http://hdl.handle.net/20.500.11752/ILC-567.

Frontini, F., Gamba, F., Monachini, M., & Broeder, D. (2021b) *SSHOC Multilingual Metadata*, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa. Available at http://hdl.handle.net/20.500.11752/ILC-568.

IATE (Interactive Terminology for Europe). https://iate.europa.eu/home

Institut de l'information scientifique et technique (Inist) - CNRS/UAR76. (2021). Loterre Open Science Thesaurus. https://dx.doi.org/10.13143/lotr.9297.

Web references

Butterfield, A., Ngondi, G. E., & Kerr, A. (2016) (Eds.) *A Dictionary of Computer Science* (7th ed.). Oxford, UK: Oxford University Press. https://www.oxfordreference.com/display/10.1093/acref/9780199688975.001.0001/acref-9780199688975.

CLARIN. https://www.clarin.eu/.

DARIAH-EU. https://www.dariah.eu/

DHA (Digital Humanities Austria). https://digital-humanities.at/en

Faloppa, F. (2010) Hapax. In *Enciclopedia dell'Italiano*. Available at https://www.treccani.it/enciclopedia/hapax_(Enciclopedia-dell%27Italiano)/.

FOSTER Open Science. https://www.fosteropenscience.eu/about#download

GO FAIR. https://www.go-fair.org/fair-principles/

Ince, D. (2019) (Ed.) *A Dictionary of the Internet* (4th ed.). Oxford, UK: Oxford University Press. https://www.oxfordreference.com/display/10.1093/acref/9780191884276.001.0001/acref-9780191884276.

Medical Subject Headings (MeSH). https://www.nlm.nih.gov/mesh/meshhome.html

Simons, P. M. (2015) Ontology. In *Encyclopaedia Britannica*. Available at https://www.britannica.com/topic/ontology-metaphysics.

TaDiRAH. https://tadirah.info/

Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/index.html

Appendix

	Concept	**Subclass**
1	ARCHE	Instance
2	B-Centre	CLARIN Structure
3	Board of Directors	CLARIN Governance
4	C-Centre	CLARIN Structure
5	CCR	CLARIN Core Services
6	CKCMC	Instance
7	CLARIN	Top node
8	CLARIN Café	CLARIN Initiatives and Events
9	CLARIN centre	CLARIN Structure
10	CLARIN Newsflash	CLARIN Initiatives and Events
11	CLARIN Resource Families	CLARIN Initiatives and Events
12	CLARIN workshop	CLARIN Initiatives and Events
13	CLIC	CLARIN Governance
14	CMDI	CLARIN Core Services
15	CMDI component registry	CLARIN Core Services
16	CMDI component	CLARIN Core Services
17	CMDI profile	CLARIN Core Services
18	CMDI standard	CLARIN Core Services
19	content search	CLARIN Core Services
20	DHCR	CLARIN Jointly Maintained Services
21	ERIC	CLARIN Governance
22	Europeana	Instance
23	K-Centre	CLARIN Structure
24	KonText	Instance
25	LINDAT/CLARIAH-CZ	Instance
26	LRS (Language resource Switchboard)	CLARIN Core Services
27	national coordinator	CLARIN Governance
28	Oxford Text Archive	Instance
29	Tour de CLARIN	CLARIN Initiatives and Events
30	VCR	CLARIN Core Services
31	VLO	CLARIN Core Services
32	WordNet	Instance

Table A. List of ‘CLARIN core’ concepts and subclass they belong to

	Concept	**Subclass**
1	data analysis	Research Data Management
2	data collection	Research Data Management
3	data curation	Research Data Management
4	data infrastructure	Research Data Management
5	data lifecycle	Research Data Management
6	data processing	Research Data Management
7	data processing on demand	Research Data Management
8	data subject	Type of Data
9	database	Research Data Management
10	depositing	Research Data Management
11	digitisation	Research Data Management
12	EOSC	Instance
13	FAIR data	Type of Data Open Science
14	FAIR principles	Type of Data Open Science
15	GDPR	Open Science
16	information science	Disciplines
17	IPR	Open science
18	Interdisciplinary research	Research Infrastructure
19	interoperability	Type of Data Open Science
20	language data	Type of Data
21	language data infrastructure	Research Data Management
22	licence	Open Science
23	linked data	Type of Data
24	linked open data	Type of Data Open Science
25	metadata	Type of Data
26	open access	Open Science
27	open data	Type of Data Open Science
28	open science	Research Data Ecosystem
29	open source	Type of Data Open Science
30	PID	Open Science
31	research community	Research Infrastructure
32	research data management	Research Data Ecosystem

33	research data repository	Research Infrastructure
34	research infrastructure	Research Data Ecosystem
35	SAML metadata	Type of Data
36	semantic interoperability	Open Science
37	semantic web	Open science
38	SSH	Disciplines
39	SSH Open Marketplace	CLARIN jointly maintained services→ SSHOC European RI Landscape
40	UI	CLARIN Initiatives and Events
41	web service	Open science

Table B. List of 'Open Science' concepts and subclass they belong to

	Concept	**Subclass**
1	AI	Disciplines
2	alignment	Techniques
3	annotation	Techniques
4	annotation tool	Tools
5	annotator	Techniques
6	ASR	Techniques
7	automatic annotation	Techniques
8	CMC	Resources
9	collocation	Disciplines
10	computational linguist	Practitioners
11	computational linguistics	Disciplines
12	computational tool	Tools
13	concept	Resources
14	concordancer	Tools
15	corpora	Resources
16	corpus linguistics	Disciplines
17	corpus querying	Techniques
18	deep learning	Disciplines
19	dependency parsing	Techniques
20	DH	Disciplines
21	digital text	Type of Data
22	HLT	Disciplines
23	ICT	Tools
24	information extraction	Techniques
25	language data	Type of Data
26	lemmatization	Techniques

27	lemmatizer	Tools
28	lexical resource	Resources
29	lexicon	Resources
30	linguist	Practitioners
31	linguistic analysis	Techniques
32	linguistic annotation	Techniques
33	language resource	Synonym of Resources
34	machine learning	Disciplines
35	machine translation	Disciplines
36	manual annotation	Techniques
37	multimodal corpora	Resources
38	named entity	Techniques
39	natural language processing pipeline	Disciplines
40	natural language processing tool	Tools
41	NER	Techniques
42	NLP	Disciplines
43	OCR	Techniques
44	ontology	Resources
45	oral archive	Resources
46	oral history	Resources
47	parallel corpora	Resources
48	parser	Tools
49	parsing	Techniques
50	plain text	Type of Data
51	POS tagging	Techniques
52	processing pipeline	Techniques
53	query	Techniques
54	query syntax	Techniques
55	semantics	Disciplines
56	sentiment analysis	Techniques
57	stemming	Techniques
58	syntactic parsing	Techniques
59	tag	Tools
60	tagger	Tools
61	tagging	Techniques
62	tagset	Tools
63	TEI	Standards
64	term extraction	Techniques
65	text analysis	Techniques
66	text corpus	Resources

67	text mining	Techniques
68	text normalisation	Techniques
69	text processing	Techniques
70	textual data	Type of Data
71	tokenisation	Techniques
72	topic modelling	Techniques
73	training materials	Resources
74	translation studies	Disciplines
75	treebank	Resources
76	UD	Techniques
77	virtual collection	Resources
78	workflow	Research Data Management
79	WSD	Techniques

Table C. List of 'LRT' concepts and subclass they belong to

Term	**Parent concepts**
Corpora	CLARIN Resource Families Resources
Tools	Language Resources and Technologies CLARIN Resource Families
FAIR data	Type of Data Open Science Guidelines
Research Data Management	Research Data Ecosystem Research Infrastructure
Open data	Type of Data Open Science
Linked open data	Linked data Open data
Lexical Resources	CLARIN Resource Families Resources

Table D. Concepts with multiple parent concepts

Concept	**Related Concepts**
annotation tool	annotation
CKCMC	CMC
	K-Centre
CLARIN	ERIC
computational linguistics	computational linguist
	computational tool

corpus linguistics	corpora
lemmatizer	lemmatisation
parser	parsing
tagger	tagging
text analysis	machine learning
VCR	virtual collection
WSD	NLP
NLP	Techniques
	Tools
national coordinator	national consortium
processing tool	text processing
DHCR	DH
SSHOC	SSH
interoperability	semantic interoperability
interdisciplinary research	disciplines
research community	practitioners
digital text	corpora
General Assembly	Scientific Advisory Board
	Board of Directors
National Coordinators' Forum	Technical Centres Committee
Board of Directors	National Coordinators' Forum
	Technical Centres Committee
	Thematic Committees
Central Hub	Technical Centres Committee
	Scientific Advisory Board
	Board of Directors
	National Coordinators' Forum
	Thematic Committees
	Standing Committee for CLARIN Technical Centres
Semantic Web	Linked Data

Table E. Associative relationships among concepts