

# MASTERARBEIT | MASTER'S THESIS

Titel | Title

FRED as an Automatic Term and Relation Extraction Tool: A Comparative Study  
of Machine-Generated Ontologies and Human-Designed Terminological  
Systems in the Medical Domain

verfasst von | submitted by

Ilaria Granzini

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of  
Master of Arts (MA)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt |  
Degree programme code as it appears on the  
student record sheet:

UA 070 331 348

Studienrichtung lt. Studienblatt | Degree  
programme as it appears on the student  
record sheet:

Masterstudium Translation Deutsch Italienisch

Betreut von | Supervisor:

Assoz. Prof. Mag. Dr. Dagmar Gromann BSc



## Acknowledgments

*A mia madre ai miei fratelli, senza il cui sostegno non avrei avuto il coraggio di perseguire e raggiungere molti degli obiettivi che ho realizzato nella vita.*

*A Rudy, con cui condivido i miei progetti più importanti e il futuro che stiamo costruendo insieme.*

*A Diletta e Veronica, che nonostante negli anni abbiamo percorso nostre strade diverse, sono sempre rimaste accanto a me, ad ogni passo.*

*Nicht zuletzt möchte ich meiner Betreuerin, Frau Prof. Dr. Gromann, für ihre fachliche Unterstützung, ihre konstruktiven Anmerkungen und ihre Geduld während der Entstehung dieser Arbeit danken.*

## Table of Contents

|                                                  |    |
|--------------------------------------------------|----|
| 1. Introduction .....                            | 1  |
| 2. Theoretical Background .....                  | 3  |
| 2.1 Terminology .....                            | 3  |
| 2.1.1 Onomasiological Approach .....             | 5  |
| 2.1.2 Concepts are Clear-cut .....               | 7  |
| 2.1.3 Definitions must be Intensional .....      | 8  |
| 2.1.3.1 Logical Concept Systems .....            | 9  |
| 2.1.3.2. Ontological Concept Systems .....       | 12 |
| 2.1.4 Univocity Principle .....                  | 13 |
| 2.1.5 Synchronic Study of Languages .....        | 15 |
| 2.2 Ontologies.....                              | 15 |
| 2.2.1 Intensional Relation .....                 | 16 |
| 2.2.2 Ontological Commitment.....                | 18 |
| 2.2.3 Intended Models.....                       | 19 |
| 2.2.4 Ontology as Computational Artefact .....   | 20 |
| 2.3 Semantic Web and Ontology Languages.....     | 23 |
| 2.4 Ontology Learning.....                       | 26 |
| 2.4.1. Ontology Learning Techniques.....         | 29 |
| 2.4.1.1. Linguistic Techniques .....             | 30 |
| 2.4.1.2. Statistical Techniques.....             | 32 |
| 2.5 FRED: Machine Reading for Semantic Web ..... | 35 |
| 2.5.1 Software Structure .....                   | 36 |
| 2.5.2 Tasks and Functional Properties .....      | 37 |
| 2.5.3 Results of FRED .....                      | 39 |
| 2.5.4 Theoretical Foundations of FRED .....      | 41 |
| 2.5.5 Translation from DRT to OWL .....          | 42 |

|                                                         |     |
|---------------------------------------------------------|-----|
| 2.6 Ontologies and Terminologies in Comparison.....     | 43  |
| 2.6.1 Syntax .....                                      | 45  |
| 2.6.2 Semantic.....                                     | 48  |
| 2.6.3 Pragmatics.....                                   | 49  |
| 2.6.4 Conclusions and Resulting Implications.....       | 52  |
| 3. Related Work.....                                    | 54  |
| 4. Research Methodology.....                            | 58  |
| 4.1 Production of Manually Created Concept System ..... | 59  |
| 4.2 Editing of TBX File in SDL Multiterm.....           | 60  |
| 4.3 Online Semantic Analysis with FRED .....            | 63  |
| 4.4 Comparison Method .....                             | 64  |
| 5. Results .....                                        | 69  |
| 5.1 Quantitative Findings .....                         | 69  |
| 5.2 Qualitative Findings .....                          | 74  |
| 6. Discussion .....                                     | 81  |
| 7. Conclusion.....                                      | 85  |
| 8. Bibliography.....                                    | 87  |
| 9. Appendix .....                                       | 100 |
| 9.1 Text Sources .....                                  | 100 |
| 9.2 Parsed Results in Python .....                      | 107 |
| Abstract (English) .....                                | 121 |
| Abstract (German).....                                  | 122 |
| Declaration on the Use of AI-Enabled Tools .....        | 123 |

## List of Figures

|                                                                                                                                                                                                                          |    |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figure 1: Semiotic triangle (Odgen & Richards 1923: 11) .....                                                                                                                                                            | 3  |
| Figure 2: Semiotic Triangle.....                                                                                                                                                                                         | 6  |
| Figure 3: Extension of semiotic triangle (Suonuuti 1997: 13) .....                                                                                                                                                       | 7  |
| Figure 4: Logical relation exemplified in biological taxonomy.....                                                                                                                                                       | 10 |
| Figure 5: Logical concept systems (Nuopponen 2018: 459).....                                                                                                                                                             | 11 |
| Figure 6: Ontological concept relations (Nuopponen 2018: 462).....                                                                                                                                                       | 13 |
| Figure 7: Relationship between phenomena in reality, their perception, their<br>conceptualization, the languages used to describe them, the intended models and<br>the resulting ontology (Guarino et al. 2009: 9) ..... | 18 |
| Figure 8: Spectrum of language L (Uschold 2004).....                                                                                                                                                                     | 21 |
| Figure 9: Ontology classification by degree of formalization (adapted from<br>Gromann 2015:24) .....                                                                                                                     | 22 |
| Figure 10: Semiotic triangle under an ontological perspective (Guarino et al. 2009:<br>16).....                                                                                                                          | 23 |
| Figure 11: Semantic Web structure (Semantic Web Stack 2014).....                                                                                                                                                         | 24 |
| Figure 12: Ontology learning architecture (Cimiano et al. 2009: 247).....                                                                                                                                                | 27 |
| Figure 13: Ontology learning layer cake structure (Cimiano et al. 2009: 250) .....                                                                                                                                       | 28 |
| Figure 14: Ontology learning techniques (Asim et al. 2018: 4).....                                                                                                                                                       | 29 |
| Figure 15: FRED Software Architecture (Gangemi et al. 2013a).....                                                                                                                                                        | 37 |
| Figure 16: Comparison of Machine Reading Tools (Gangemi 2013: 360).....                                                                                                                                                  | 38 |
| Figure 17: List of prefixes used in FRED diagrams and inline examples (Gangemi<br>et al. 2017: 4).....                                                                                                                   | 40 |
| Figure 18: FRED's output (Gangemi et al. 2017: 3).....                                                                                                                                                                   | 41 |
| Figure 19: SDL Multiterm: Entry Structure Definition .....                                                                                                                                                               | 62 |
| Figure 20: SDL Multiterm: establishing of concept relations between entries.....                                                                                                                                         | 63 |
| Figure 21: FRED: settings.....                                                                                                                                                                                           | 64 |

## **List of Tables**

|                                                                                                            |    |
|------------------------------------------------------------------------------------------------------------|----|
| Table 1: Main Translation Rules from DRT to OWL (adapted from Presutti et al. 2012: 123) .....             | 42 |
| Table 2: Criteria for Semiotic Comparison of Terminologies and Ontologies (adapted from Gromann 2015)..... | 45 |
| Table 3: TBX and OWL in comparison (Gromann, 2015: 108) .....                                              | 45 |
| Table 4: Terminology and ontology processes compared (Gromann 2015: 113)....                               | 50 |
| Table 5: Summarized comparison of terminologies and ontologies. ....                                       | 52 |
| Table 6: Concept relations employed in the study .....                                                     | 60 |
| Table 7: Comparison between extracted terms in FRED and terms in SDL.....                                  | 70 |
| Table 8: Comparison between analysed relations by FRED and manually reported relations in SDL .....        | 72 |
| Table 9: Summary of Extraction Performance.....                                                            | 81 |

# 1. Introduction

*“...no language has ever possessed but a small fraction of the varied excellences of human speech, and each language represents a different vision of life”*

(J.R.R. Tolkien, *Philology: General Works*)

The focus of the present master’s thesis is the analysis about the feasibility of a terminological approach to a formal knowledge organization system, namely an ontology, automatically generated by ontology learning tool performing a term extraction from a given text. The machine reading software selected for this study is FRED (Gangemi et al. 2013b), a Semantic Web-based tool, available online.

The present study focuses on the comparison between ontologies and terminologies, thus between the analysis conducted by an ontology learning tool and a manually curated terminological concept system. Large space will be given to the differences between ontologies and terminologies: this comparison is crucial for designing the comparison model between the two approaches. Equally, FRED will be analysed in-depth, as applied theory, as machine reading tool for Semantic Web and, in future, as term-extractor.

The starting point for this research is the hypothesis, that ontology learning tools can be used as fully automatic term-extraction tool in perspective of compilation of a terminology, since it generates formal knowledge representations systems from texts.

The research questions are formulated as follows:

**RQ1:** When comparing a fully automatically generated ontology from text and a human manually curated terminological concept system, which differences can be found?

**RQ2:** Which advantages and disadvantages has an ontology learning tool as automatic term extractor?

The major objective of this research is the application of a machine reading tool based on Semantic Web technologies as term extractor. Terminology is one important building block not only of translation, but also of artificial intelligence, computational linguistic and the Semantic Web (Bassey 2005) and its development spread out during the last 20-30 years. Drouin et al. (2018: 2) write that „the task of filtering [the designations] should be made not in terms of the

relevance to texts but in relation to the conceptual structure of the domain, which naturally leads us to the issue of coherence of terms as a set”, meaning that, not only a rule-based method, i.e. use of rules that represent the knowledge learned by the system, or statistics can be satisfying for an automatic term extraction.

Resulting from the theoretical findings, the differences between ontologies and terminologies will be defined, in order to design a comparison model between a manually curated terminological concept system (epistemic knowledge) and an automatic generated concept system in form of ontology (ontological knowledge), which are empirically tested in medical domain. Therefore, the two objectives of this research are:

1. Design of a comparison model between a manually terminological concept system and an automatic generated concept system, considering the differences between natural language and ontology (formal) languages.
2. Practical application of the comparison model in the medical domain, in order to determine the adequacy of an ontology learning tool to extract a terminological concept system.

## 2. Theoretical Background

### 2.1 Terminology

De Saussure, father of the modern structuralist linguistics, laid the groundwork for terminology, years before it gained the status of science. In 1916, Saussure's lectures were posthumously edited and published by his students, who had attended his courses at the University of Geneva during the academic years 1906–1907, 1908–1909, and 1910–1911 (Prampolini 2013). During these years he delineated the two principles of structuralism and the fundamentals of semiotic and semiology. According to the first principle, signs have an arbitrary nature, i.e. the signifier denotes a precise signified unmotivated, conventionally and independently.

The signified refers to the concept, whereas the signifier refers to the sound-image, i.e. “the natural representation of the word as a fact of potential language, outside any actual use of it in speaking” (De Saussure et al. 2011). Thus, the sign represents the indissoluble union of signified and signifier. According to the second principle, the signifier has a linear nature, since it exists only in the auditory dimension. Signifiers are bound to the existing language and its community, emphasising that language is the results of previous generation and it changes as the time passes by (mutability of signs). He also introduced the dichotomy *langue-parole*, which represents the bond established between the system of the language (*langue*) and the linguistic act, the concrete realization of langue (*parole*) (De Saussure et al. 2011).

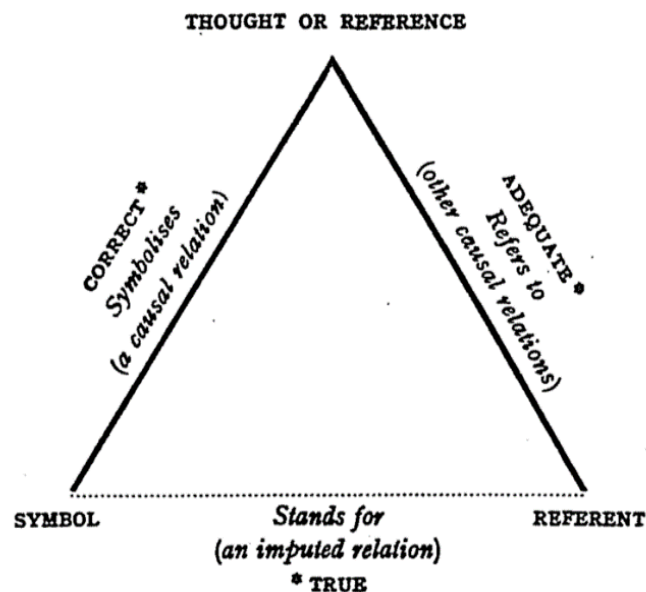


Figure 1: Semiotic triangle (Odgen & Richards 1923: 11)

The relation between signifier, signified and referent is one of the cornerstones of terminology as understood nowadays: the referent and the signified are linked by signifier, that can be expressed by sound-image, sign, symbol or ideogram. This theory was investigated deeper by Odgen & Richards (1923) and named as *triangle of reference*, or *triangle of meaning* or *semiotic triangle* (Figure 1). As thought of reference is intended the *signified*, i.e. the concept, while the symbol is the *signifier* and the *referent* represents the object. This formula is currently in use in sciences, such as semiotics and terminology: this bond is applied in different domains, although the designations at the edges on the triangle may change.

Terminology science, or terminology studies, is defined by ISO 1087 (2019: 3.1.12) as “science studying terminologies, aspects of terminology work, the resulting terminology resources, and terminological data”, whose scope is the terms’ standardization and organization of a domain of interest. A terminology is a set of designations and concepts proper of a domain of interest or subject, while terminology work is the terminology managing of collection, description, processing and presentation of concepts and their designations.

Eugen Wüster, Austrian multi-faceted engineer, industrialist and terminologist, draws inspiration from structuralism and established the terminology theory. As a young student, his interest in Esperanto and lexicography developed further in what is known today as Terminology Science and the Technical Committee for Terminology Standardization of the International Organization for Standardization (ISO/TC 37). Wüster’s General Theory of Terminology (GTT) is a practical, normative and prescriptive approach, which is described in Wüster (1979). Its core is based on concepts and their delimitation to each other by definitions. At first, GTT was accepted and implemented, and its followers are grouped in what is known as Vienna’s School of Terminology: a practical and concept-oriented approach, which organizes terminological information by means of concepts, the basic units of meaning.

Since the last years of the 20<sup>th</sup> century GTT has been discussed by different theories of Terminology, mostly by Socio-Cognitive Theory by Temmermann (2000) and Frame Based Terminology by Faber (2012). Temmermann (2000) focuses on the critical principles of GTT, analysing, criticizing and overturning the five milestones of GGT. The basis for her analysis is the “excessive concern with standardisation” (Temmermann 2000: 2), due the fact that the standardization limits the “possibilities for unprejudiced research in the field of Terminology” (Temmermann 2000: 3). Moreover, Cabré Castellví (2003: 171) has analysed and clarified that “traditional terminology” has been criticized under three different points of view: of cognitive

science, language sciences and communication sciences. The cognitive psychology and philosophy point out the absence of a clear separation between general and specialized knowledge, as well the importance of interlocutors in discourses and the importance of culture in the perception of reality. Language sciences, especially linguistics and sociolinguistics, stresses the rigid division between general and specialized language and the social bases of special languages, highlighting the importance of semantics and pragmatics even in these types of language sciences. Communication sciences (Cabr  Castellv  2003: 172) propose different models in the form of frames, which help the integration of specialised communications “as a set of options in a single schema, rather than treating it as a different type of communication”. Faber (2012) clarifies a cognitive approach to terminology, which connects specialized knowledge representation and cognitive linguistics and semantics. Today’s understanding of terminology science is multidisciplinary: it crosses linguistic, conceptual and communicative disciplines to provide an integrated approach to terminological entries, in order to use them correctly in professional and specialized communication (Drewer et al. 2017).

The General Theory of Terminology is built upon five principles. In this section each principle will be analysed in depth, and extended as today’s understanding of terminology science, in order to clarify the fundamental elements in a structured way.

### **2.1.1 Onomasiological Approach**

At the core of GTT is the concept and its definition, i.e. the onomasiological study of language with a focus on the meaning of the sign. Concept is defined as “unit of knowledge created by a unique combination of characteristics” (ISO 1087 2019: 3.2.1). A concept is processed by ‘abstraction’, a mental process, which describes an object by its characteristics. These characteristics combined create the concept itself. Temmermann (2000: 5f.) criticizes that the Vienna’s School of Terminology refers to the concept as a part of the world outside the language, reducing the mental activity of concepts’ naming to a mental activity. Indeed, Felber (1984) states that elements or sound patterns become terms only when they are associated with concepts.

Concepts are distinguished into general and individual: general concepts are concepts that can describe similar objects with common properties (e.g. *planet*), while individual concepts refer to a single object (e.g. *Mars*) (ISO 1087 2019). An object is whatever is perceivable or conceivable and is defined by its properties. Therefore, characteristics, which describe a concept, are abstractions of properties, which define the object itself (Suonuuti, 1997: 12).

In Klein and Smith (2010: 435f.) the three different interpretations of concepts are listed: as universal, individual and general terms in controlled language. Universal concepts are not properly correct, because concepts represent the epistemic knowledge about an object, not the real-world object. Individual concepts are controversial, because the mental representation of an object denotes an individual representation and terminologies should represent agreed and shared knowledge. General terms in controlled language cannot replace the representation of the signified in natural, artificial language and neither in ontologies. Further, Klein & Smith (2010) recommend identifying the concept as representation of agreed meaning of designations, because it exists only within the concept system, while Nuopponen (2010) defines the concept as onomasiological by structuring the specialized language, i.e. it has to be agreed upon by experts.

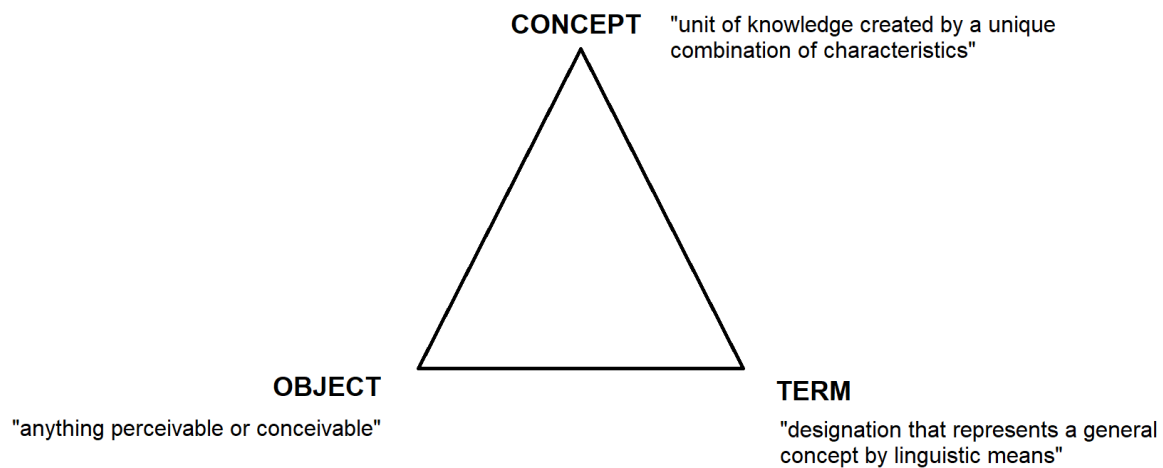


Figure 2: Semiotic Triangle

Concepts are not independent entities, they exist only in relations with objects, definitions and terms, to which they refer. Thus, concepts are mental phenomena, while objects, definitions and terms refer to the physical world (Suonuuti 1997). Figure 2 represents graphically the relation between object, concept and terms, building the semiotic triangle (Odgen & Richards 1923). In Figure 3, the relations between concepts, terms, definition and objects are represented as extensions of the semiotic triangle (Suonuuti 1997: 13).

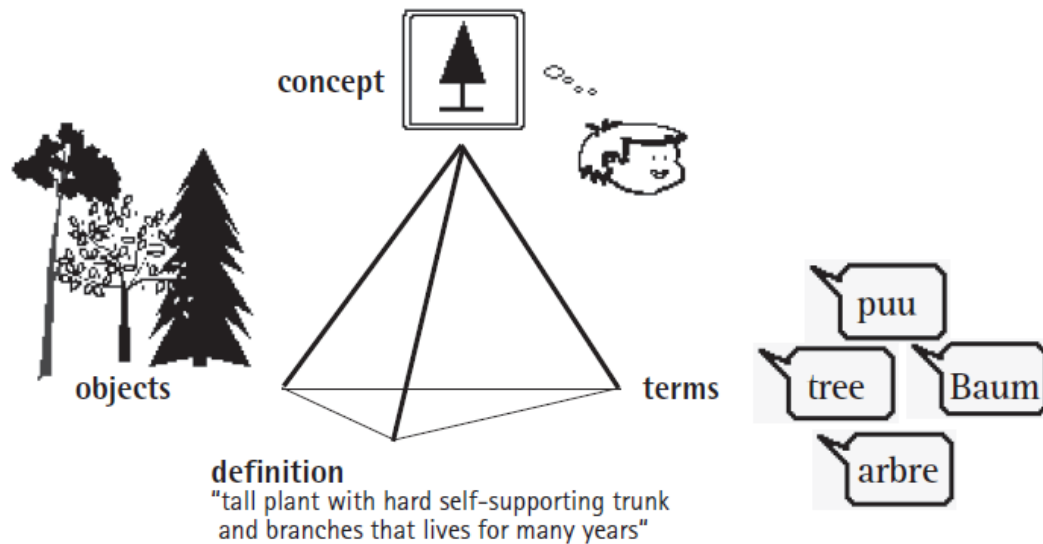


Figure 3: Extension of semiotic triangle (Suonuuti 1997: 13)

In terminology science characteristics are part of natural language definitions: therefore, definitions depend on their positioning in the concept systems (Gromann 2015). Each characteristic can be categorized in a specific concept relation, positioning the concept within the concept system. Characteristics can be of two types: essential characteristics, which can be delimiting characteristics, therefore they are used for defining the concept by distinguishing it from its related concepts (ISO 1087 2019); and non-essential characteristics. Wright (1997) describes two types of characteristics, according to the relation between the characteristic itself and the concept: attributes, which are inherent to the concept and does not imply any relation to other concepts; and relations, which implements information about other concepts, that are correlated to the analysed concept.

### 2.1.2 Concepts are Clear-cut

Concepts can be defined only in relation to other concepts (Nuopponen 2018), namely in a concept system, which is created by studying the characteristics of the concepts themselves. In fact, characteristics are elements of concepts, which are needed to describe or identify a certain quality of individual objects (Felber 1984: 115). Concept systems are “sets of concepts structured in one or more related domains according to the concept relations among its concepts” (ISO 1087 2019: 3.2.28). They are fundamental for terminological concept analysis and terminology work.

According to Felber (1984: 120) concepts have to be studied in a system, due the fact that they are composed by characteristics and interrelated because of the same characteristics, which

means that the relations between concepts are based on their intention. In order to place a concept in a concept system it is necessary to delineate the concept itself, and this can be made only by comparison of characteristics of concepts (Temmermann 2000: 7f.). Hence, GTT is mostly interested in the concept relationships, logical and ontological. So as to ensure an effective and efficient communication and to avoid unambiguousness, the concepts have to be clearly delineated, i.e. clear-cut. The criticism moved by Temmermann (2000) against the second principle of GTT is based on the fact that many concepts are not clear-cut, because sets of characteristics are flexible and they are not restricted and it is impossible to give “a definition of necessary and sufficient characteristics of a category” (Temmermann 2000: 8). Temmermann (2000) concludes her analysis declaring that flexibility is necessary for categorisation and communication, because they lead to progress in understanding.

### **2.1.3 Definitions must be Intensional**

Concepts can be also identified by their intension and their extension. Intension is the set of essential and delimiting characteristics that define a concept (ISO 1087 2009: 3.2.6), while extension determines the set of objects, to which the concept refers (Suonuuti 1997: 13). Terminological definitions can be of three types: (a) intensional, (b) extensional and (c) part-of-whole.

- (a) intensional definitions reflect the Aristotelian *genus-differentiae*: this kind of definition describes genus (type) as a superordinate concept, the designation, and differentiae as the delimiting characteristics in natural language (Gromann 2015). In natural language, characteristics are represented in the intensional definition.
- (b) extensional definitions: all objects belonging to the superordinate concept are listed in order to define it (ISO 704 2009), i.e. the enumeration of the concept’s extension (Temmerman 2000).
- (c) part-of-whole or partitive definitions: all parts of the composite concept are listed (Temmerman 2000).

Definitions must be intensional, in order to place concepts in a concept system. A concept system “reflects the structured organization of a single well-defined domain” (De Bessé 1997: 65f.). Concept formation depends both on the identification of characteristics and on the concepts’ relations to each other. Pozzi (2001) describes two limits of intensional relations: firstly, in some cases not all characteristics referred to a concept might be known; secondly, terminological concept systems are hierarchies in which the superordinate concept has not its

own superordinate concept from which it obtains its characteristics. In addition, Seppälä (2012) describes a further limit: the exact nature of the characteristics is not defined; thus, it is the terminologists' responsibility to decide and define the concept system.

ISO 1087 (2019) defines three different types of concept relations: (a) generic, (b) partitive and (c) associative. Generic and partitive relations both represent hierarchical relations (ISO 1087 2019).

- (a) Generic relation is a concept relation between a generic concept and a specific concept (ISO 1087 2019): the two concepts share an identical set of characteristics, but the subordinate concept has in addition one essential delimiting characteristic (Nuopponen 2018).
- (b) Partitive relation defines a concept relation between a comprehensive concept (concept constituted of various parts) and a partitive concept (part of whole) (ISO 1087 2019).
- (c) Associative relation describes a non-hierarchical concept relation, which can be subdivided in sequential (concepts ordered by ordering criterion), spatial (based on the criterion of relative location in space), temporal (based on before-after relation) and casual relations (based on cause-effect criteria) (ISO 1087 2019).

Wüster (1975) divided the concept relations into logical relations, which originates in logic and logical reasoning, and ontological relations, which are taken from ontologies. Ontological relations represent the real relationship between physical objects (ontic relations), in fact they are based not on logical relation (Nuopponen 2018). Therefore, logical relations characterize logical concept systems, while ontological relations characterize ontological concept systems. (a) generic concept relations correspond to Wüster's logical relations, representing hierarchies, typologies and taxonomies. (b) partitive relations and (c) associative relations belong to Wüster's ontological concepts relations. As follows, the two notions of logical concept systems and ontological concept systems are described as in Nuopponen (2018), who extends and modifies Wüster's classification of logical and ontological concept relations.

#### *2.1.3.1 Logical Concept Systems*

Logical concept systems are organized by logical relations, i.e. hierarchical relations in the logical concept systems: the most abstract concept is positioned on the top and the most concrete at the bottom. A self-evident and easy example of this kind of concept systems is the

taxonomy in biology science (the organization and classification of animals), exemplified in the following schema:

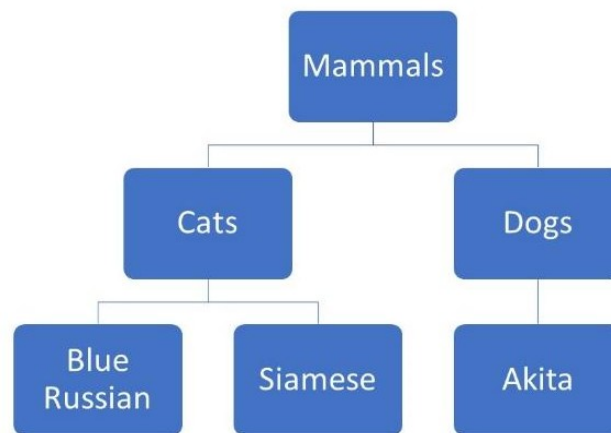


Figure 4: Logical relation exemplified in biological taxonomy

Figure 4 exemplifies a taxonomy, in which the most concrete concepts at the bottom (*Blue Russian*, *Siamese* and *Akita*) represents individuals belonging to a specific class or *specie*, while the fewer concrete concept (*Mammals*) is positioned on the top of the tree-structure of the concept system. By the analysis of this example, it is easy to assume that it is also necessary to determine the division criteria to organize the different characteristics. Hence, it is possible to identify three types of logical or generic relations between concepts according to their degree of abstraction, crucial process by writing definitions:

- (a) Logical/generic superordination: superordinate concepts have a major extension to their subordinate concepts. By writing definitions, in fact, it is important to detect the superordinate concept to the concept to be defined.
- (b) Subordination: subordinate concepts have a larger intension than superordinate concepts, i.e. they have at least one characteristic more than the superordinate concept.
- (c) Logical/generic coordination: it represents the relations between concepts on the same level of abstraction, i.e. the relations between different subordinate concepts to the same superordinate concepts (in the above example the relation between *Siamese* and *Blue Russian*). By writing definitions, it is also essential to define the coordinate concepts in order to delimitate the concepts to each other, by finding the essential and delimiting characteristics.

Logical concept relations can also be defined according to similarity of intension; in this case we can distinguish four different types of logical concept relations (Nuopponen 1994), illustrated in Figure 5.

- (a) Intensional inclusion: one of the concepts (the subordinate) includes all the characteristics of the other (the superordinate) with additional ones.
- (b) Intensional overlapping: concepts have common characteristics, but they both have some additional ones. The concepts can also refer to overlapping objects of reference.
- (c) Intensional disjunction: the concepts are on the same level of abstraction and exclude each other.
- (d) Intensional identity: the considered concepts have identical intension, i.e. they have the same characteristics. Intensional identity establishes the synonymy relation between two different concepts (See 2.1.4 Univocity principle).

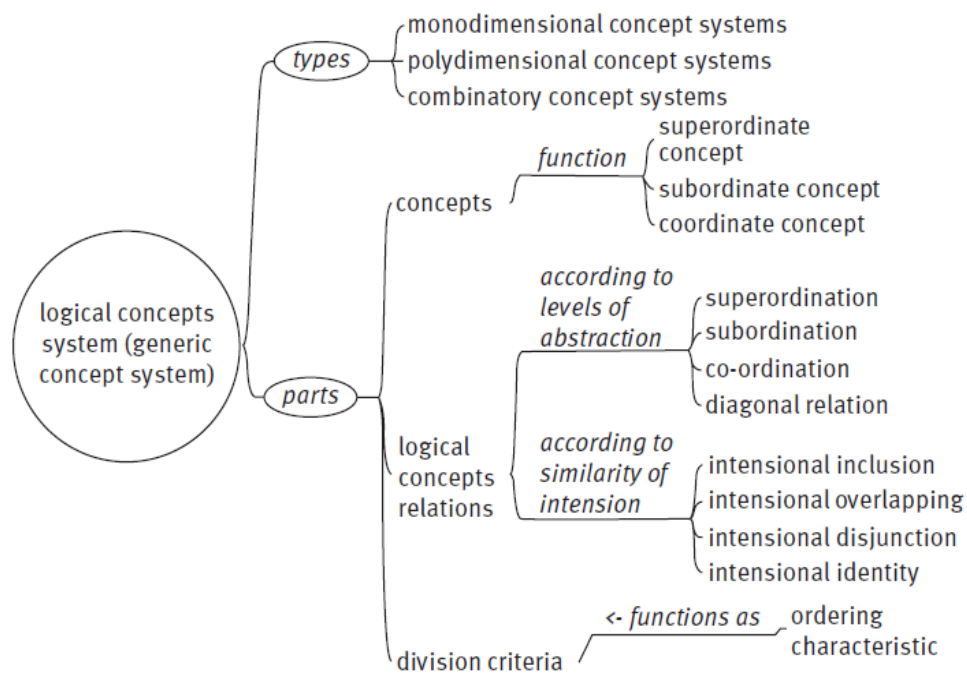


Figure 5: Logical concept systems (Nuopponen 2018: 459)

Moreover, Nuopponen (2018) adds the intensional negation as concept relation. But this case does not appear in Figure 5, because it represents intrinsically a paradox: the considered concept has opposite characteristics of another concept's intension.

The logical concept systems are polydimensional, when the referring objects can be categorized into classes according to different criteria, and combinatory, when these criteria are applied at once.

#### *2.1.3.2. Ontological Concept Systems*

Ontological concept systems are described in Wüster (1974) in order to classify those concepts relations, which are based on different criteria than logical relations. Ontological relations root from ontic relations, which describe relationships between individuals in the real-world. In Wüster (1974, 1975) ontological relations are divided in contiguity relations, characterized by contact in time and space, and effect relations, such as genealogic, phylogenetic, ontogenetic and descent. Nuopponen (2018) examines and broadens the different ontological concept relations, which are represented in Figure 6. These concept relations can form micro concept systems, which can be combined in macro concept systems specific to a domain or subject.

Concepts cannot be classified in one relation only, usually they are structured in mixed concept systems, combining different types of relations (Antia 2000). Another complex task stems from the possible different perspective of an object, which can induce different concepts. Barsalou (1992) proposes to introduce the theory of frames, in order to represent the variability of a concept in different contexts: frames vary values across different applications.

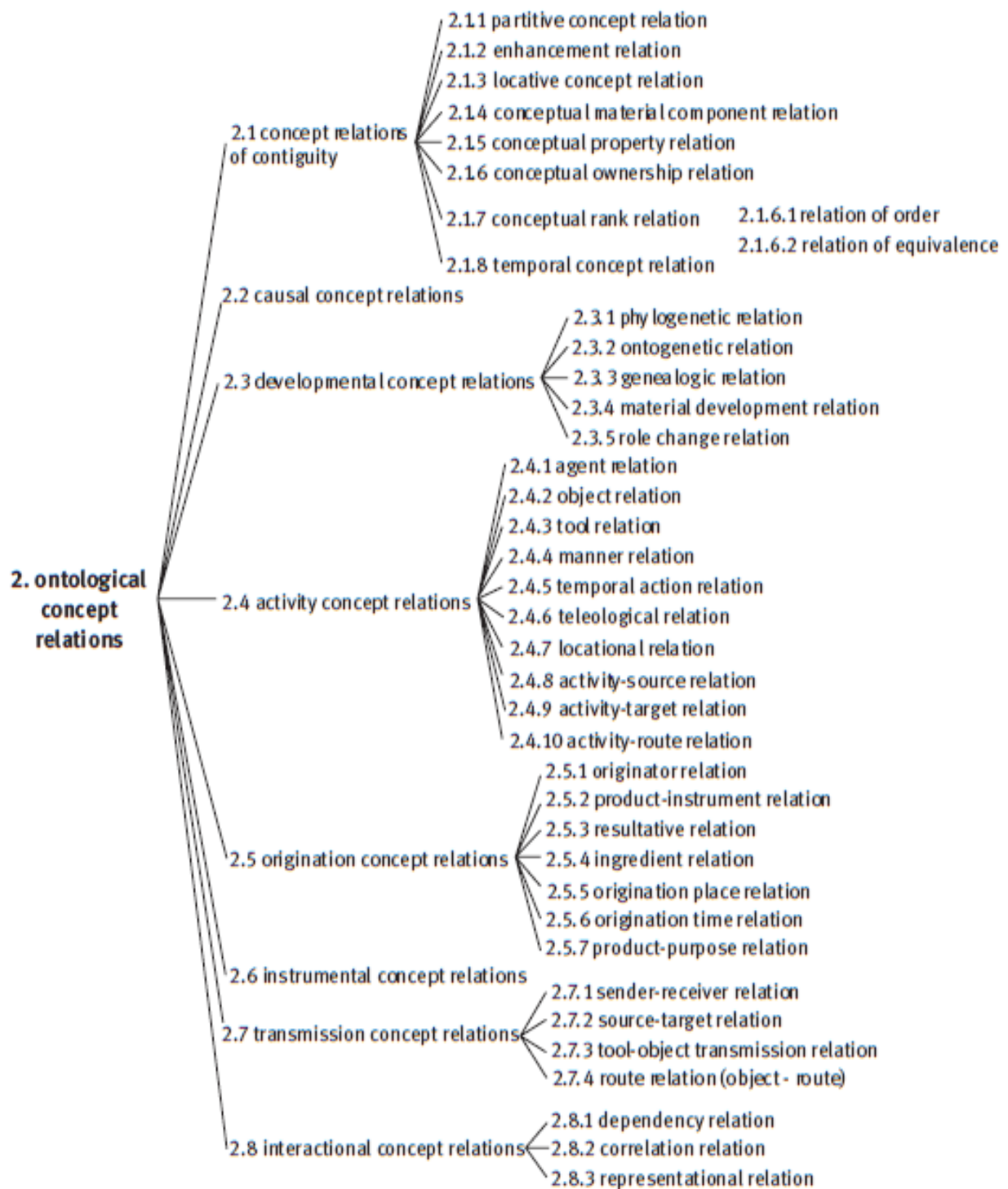


Figure 6: Ontological concept relations (Nuopponen 2018: 462)

### 2.1.4 Univocity Principle

Each concept must be described only by one designation, and each designation must describe only one concept. A designation is defined as “representation of a concept by a sign which

denotes it” (ISO 1087 2019: 3.4.1). Designations may be linguistic or non-linguistic, which can include terms, appellations, proper names or symbols. According to ISO 1087 (2019 3.4.2), a term is a “designation that represents a general concept” and can be simple, e.g. a simple-word term, or complex, e.g. a multi-word term (Arntz et al. 2009).

By introducing the univocity principle, it is required a clarification of the notion of synonymy. Synonymy is defined in ISO 1087 (2019: 3.4.15) as the “relation between designations in a given language representing the same concept”. Such designations may be interchangeable in every context; however, this last condition does not fully account for textual and contextual variability of designation use. By contrast, the relation between designations in different natural languages, which refer to the same concept is defined as equivalence.

According to Arntz et al. (2009), equivalence can be represented in four different cases. The first case is the full conceptual equivalence, in which all characteristics are identical. In this case, synonymous designations are recorded within the same terminological entry as part of terminological entry data. The second is partial overlap, which occurs when two different designations share same characteristics, but they also have additional ones. The third case is inclusion, whereby one concept incorporates another, while it potentially has additional characteristics. Both inclusion and partial overlapping are interesting cases, because they can manifest either within the terminological entry, between two different entries or between two different terminological works. Latest, zero equivalence defines the relation between two concepts without overlap of characteristics. This case is also known as conceptual or designation gap, which can be resolved through three different methods: translation from another language, creation of a new term, or description in the target language (Arntz et al. 2009) or adopting of loanwords. An example of loanword is the German word *Doppelgänger* adopted in English to describe the concept of a person who looks exactly like another person.

Mayer (1998) further describes different terms relations, including intralingual synonymy, interlingual equivalence, fuzzy intralingual synonymy and fuzzy interlingual equivalence and states that such term relations should be employed when concept relations alone are insufficient describe the degree of equivalence or synonymy.

The axiom according to which one term corresponds to one concept is now considered obsolete, since concepts are not language independent. Strict adherence to this principle would require the elimination of both synonymy and polysemy. Although contemporary

terminological work permits synonymy and equivalence, it remains preferable to assign a single preferred term to each concept (Temmermann 2000: 13), however it is recommended to avoid a certain grade of synonymy in order to establish a certain grade of standardization (Cabr  Castellvi 2003). Polysemy and synonymy are therefore functional aspects of terminology (Temmermann 2000: 14).

Term harmonization refers indeed to process of selecting designations for harmonized concepts. This process must follow concept harmonization, which aims to identify and reduce differences between overlapping or closely related concepts (ISO 1087 2019). The scope of terms is, in fact, related to the organization of similar physical objects into categories. This cognitive process, defined as categorization, is crucial for structuring of human knowledge (Budin 1996: 44).

### **2.1.5 Synchronic Study of Languages**

Terminology science is not concerned with language evolution; rather, it does not investigate the evolution and dynamics of a domain-specific language (Aussenac-Gilles et al. 2008). This onomasiological approach differs from the semasiological perspective adopted in linguistics. While terminology focuses on domain-specific terms, linguistics is interested on language as whole. Terminology adopts a synchronic perspective of languages, indeed.

According to GTT the relation one term = one concept is considered permanent and arbitrary, and it does not evolve over time. Consequently, the spatial and temporal collocation of terms is not a concern of GTT; instead, its objective is the shaping and the understanding of specialized terms. On the contrary, the Socio-Cognitive Theory of Terminology argues that categories evolve and that both concepts and their designations undergo semantic change, as the human understanding changes and evolves (Temmermann 2000: 14ff.). From this perspective, the diachronic study of terms becomes necessary for understanding concepts (Dury 2005). As well, Temmermann (1998) emphasizes that evolution has a leading role in the comprehension of units such as concepts.

## **2.2 Ontologies**

Ontology carries a dual meaning and in literature is distinguished as the uncountable noun *Ontology*, capitalized, to denote the philosophical discipline, and the countable noun *ontology*, lowercase, referring to computational artifact. The former delineates the philosophical study of

being its structure and its formalisation by classifying it in classes and their relationships (Hitzler et al. 2009). The origins of Ontology can be traced back to Aristotle, who examined the structure and nature of reality in “Metaphysics” (Guarino et al. 2009).

The latter meaning, that can be also defined as computational ontologies, concerns the “explicit specifications of conceptualizations” (Guarino et al. 2009). Over the last 20-30 years, the term ontology has been widely adopted in computer science and artificial intelligence, where it is understood as formal description of knowledge within a domain of interest, expressed through logic statements or axioms, and represented as a set of concepts, relations and meta-data (Gromann 2013). Computational ontologies classify the knowledge into classes and instances in order to render domain-specific knowledge machine-readable (Drewer et al. 2017). In this sense, computational ontologies study categories of things through explicit conceptualization, a crucial process for ontology building that involves producing an abstract and simplified representation of the physical world. At the same time, significant challenges remain regarding ontological knowledge representation, as ontologies may „fail to relate to real-world objects and describe knowledge as structured in the natural language by a specific community of speakers” (Gromann 2015: 14).

Conceptualization is defined by Genesereth & Nilsson (1987) as an extensional relational structure, represented by the tuple  $(D, \mathbf{R})$  where  $D$  is a set called the universe of discourse and  $\mathbf{R}$  is a set of relations defined on  $D$ . Accordingly, each element of  $\mathbf{R}$  represents an extensional relation, reflecting a specific state of the world and involving the elements of  $D$ . However, this formulation has been criticised for being too simplified, as the conceptualization itself should remain stable when the universe  $D$  changes. Consequently, such definition of conceptualization does not fully satisfy the main purposes of ontologies, particularly their role in providing a stable and reusable representation of knowledge across varying contexts.

### **2.2.1 Intensional Relation**

In Guarino et al. (2009: 5) intensional relations i.e. conceptual relations are depicted as functions mapping a set of maximal world states onto extensional relations. In order to provide a mathematical definition of intensional relation, the concepts of system and world within computational ontologies need clarification. As system is intended the portion of reality to be modelled, whereas world is described as an “ordered set of world states, corresponding the system’s evolution in time” (Guarino et al. 2009: 6).

Accordingly, an intensional relation, or conceptual relation, is defined as follows:

- $S$  is an arbitrary system;
- $D$  is an arbitrary set of distinguished elements of  $S$ ;
- $W$  denotes a world, or a set of possible worlds, corresponding to the set of world state  $S$ ;
- The tuple  $\langle D, W \rangle$  constitutes the domain space for  $S$ ;

The intensional relation  $p^n$  of arity  $n$  on  $\langle D, W \rangle$  is therefore defined as a total function from the set  $W$  into the set of all  $n$ -ary (extensional) relations on  $D$  (Guarino et al. 2009: 6):

$$p^n: W \rightarrow 2^{D^n}$$

A conceptualization  $C$ , or intensional relation structure, can be represented mathematically as the triple:

$$C = (D, W, R)$$

Where:

- $D$  is a universe of discourse;
- $W$  a set of possible worlds;
- $R$  a set of conceptual relations on the domain space  $\langle D, W \rangle$ .

(Guarino et al. 2009: 6f.)

In order to refer to the elements of a conceptualization, it is necessary to introduce a system of symbols, namely a language  $L$  committed to the conceptualization. Language commitment refers to the competent use of language for representing and adapting the conceptualization. Once a language  $L$  has been selected, it is necessary to admit only those models that are accepted by conceptualization. Indeed, possible models represent different extension of the conceptual relations. Therefore, the conceptualization itself, which is often implicit, must be made explicit through the specification of these extensions. In Figure 7 the process of conceptualization is illustrated.

Conceptualizations may be specified either by extension or by intension. Extensional specification consists in listing all the extensions of every conceptual relation across all possible

worlds. Although this approach is impractical, but in some, it might be useful in certain cases where conceptual relations extensionally can be illustrated through examples. Intensional specification, by contrast, is based on axioms and postulates, expressed in formal and machine-readable languages.

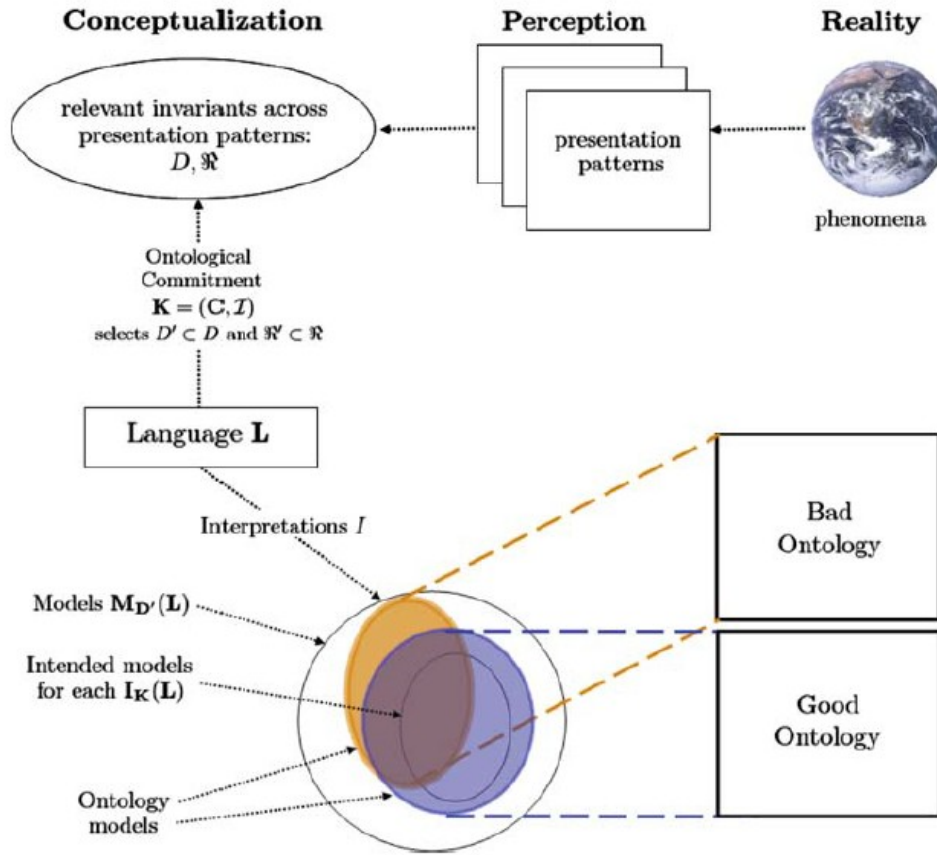


Figure 7: Relationship between phenomena in reality, their perception, their conceptualization, the languages used to describe them, the intended models and the resulting ontology (Guarino et al. 2009: 9)

### 2.2.2 Ontological Commitment

An ontological commitment, or intensional first-order structure for a language  $\mathbf{L}$  with a vocabulary  $\mathbf{V}$  and an intensional relation structure  $\mathbf{C} = (D, \mathcal{W}, \mathcal{R})$ , is defined as the tuple  $\mathbf{K} = (\mathbf{C}, \mathbf{I})$ , where  $\mathbf{I}$  denotes an intensional interpretation. This interpretation is represented as a total function mapping each vocabulary symbol in  $\mathbf{V}$  either an element of  $D$  or an intensional relation belonging to the set  $\mathcal{R}$  (Guarino et al. 2009: 10):

$$\mathbf{K} = (\mathbf{C}, \mathbf{I})$$

$$I: \mathbf{V} \rightarrow D \cup \mathbf{R}$$

Accordingly, ontological commitment can be understood as an extension of the standard notion of conceptualization, namely the transition from the extensional interpretation of meaning associated with vocabulary symbols to an intensional interpretation. In this sense, ontological commitment establishes the correspondence between the symbols of a formal language and the conceptual entities and relations that constitute the underlying conceptualization.

### 2.2.3 Intended Models

Once the intensional meaning of a vocabulary has been established, the range of admissible models becomes constrained accordingly. An intended model of a first-order logical language  $\mathbf{L}$  with a vocabulary  $\mathbf{V}$  is a model  $M = (S, I)$ , where  $S = (D, \mathbf{R})$ , and  $I$  is an extensional interpretation function mapping each vocabulary symbol either to an element of  $D$  or to an extensional relation in  $\mathbf{R}$ , according to ontological commitment  $\mathbf{K} = (C, I)$  (Guarino et al. 2009: 11). Formally, the interpretation function is defined as:

$$I: \mathbf{V} \rightarrow D \cup \mathbf{R}$$

A model  $M$  is compatible with the ontological commitment  $\mathbf{K}$  if and only if the following conditions hold:

1. For all constant symbols  $c \in \mathbf{V}$ , the interpretation assigned in the model corresponds to the interpretation established by the ontological commitment:

$$I(c) = I(c);$$

2. There exists a world  $w \in \mathcal{W}$  such that, for each predicate symbol  $v \in \mathbf{V}$  there exists; an intensional relation  $p \in \mathbf{R}$  satisfying the following conditions:

$$I(v) = p$$

$$I(v) = p(w).$$

The set of intended models of  $\mathbf{L}$  according to the ontological commitment  $\mathbf{K}$  is denoted by  $\mathbf{I}_K(\mathbf{L})$  and consists of all models of  $\mathbf{L}$  that are compatible with  $\mathbf{K}$ . (Guarino et al. 2009: 11). In this framework, intended models represent those interpretations of a formal language that

faithfully correspond to the conceptualization and ontological commitments underlying the ontology.

#### 2.2.4 Ontology as Computational Artefact

Given a conceptualization  $C$ , a logical language  $L$  with vocabulary  $V$ , and an ontological commitment  $K$ , an ontology  $O_K$  for  $C$  with vocabulary  $V$  and ontological commitment  $K$  is defined as a logical theory consisting of a set of formulas expressed in  $L$ . This definition is designed so that the set of its models approximates as closely as possible the set of intended models of  $L$  according to  $K$  (Guarino et al. 2009: 11).

A model refers to a specific assignment of abstract relational structures to the vocabulary elements; consequently, each model corresponds to a distinct pattern of relationships. A possible world, by contrast, represents a combination of actual states of affairs within a given system. The number of world states therefore depends on the selected variables, whereas the possible models depend on the vocabulary and domain. Since each model corresponds to a different relational structure, the selection of an appropriate vocabulary is essential for the development of an effective ontology and for ensuring adequate expressiveness in the axiomatization process.

Language expressiveness depends on the characteristics of the adopted representation approach for language  $L$ . Figure 8, which illustrates the spectrum of language  $L$ , distinguishes the two different approaches: informal and formal language. Informal approaches are placed at the left edge of the spectrum, which permits only terms' definitions. At the right edge, the formal approaches i.e. the formal languages are ranked. There is no defined and strict boundary between formal languages and informal languages, thus in order to choose the right language, it is decisive to find the balance between expressiveness and efficiency (Guarino et al. 2009: 13). First-order logic suits this purpose, because it is flexible and enables logical reasoning.

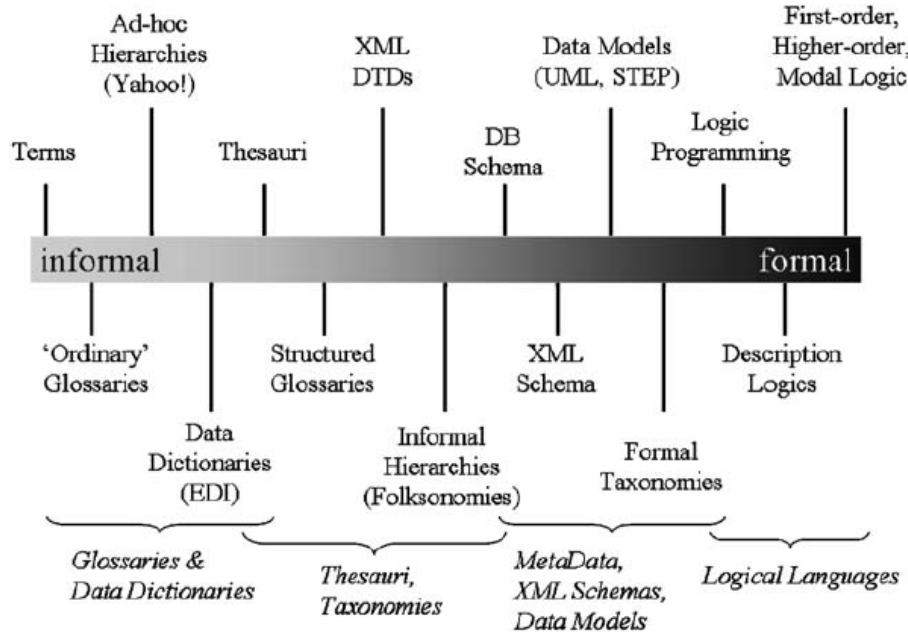


Figure 8: Spectrum of language L (Uschold 2004)

First-order logic is subdivided into two paradigms: Description Logics (DL), e.g. the Web Ontology Language (OWL), and Logic Programming (LP), e.g. Frame-Logic (F-Logic). The former constitutes a family of formal knowledge-representation languages, designed to model knowledge in a structured and machine-readable form (Baader et al. 2009: 23). The latter can be used to automatically infer the implicit knowledge from the explicitly represented knowledge.

DL represents the foundation for web ontology languages, such as OWL, which is built upon the Resource Description Framework Schema (RDFS), and its design is based on a specific subset of description logic. LP approaches, such as F-Logic, combine conceptual modelling with declarative representation style (Angele et al. 2009: 45).

Gromann (2015) proposes an ontology classification according to their degree of formalization, and states that ontologies are distinguished in first place between those dealing with conceptualization subject and those that represent reality through varying degrees of richness of structure. Figure 9 illustrates the distinction between formal and informal representation systems. The difference between these approaches is evident in the interpretation of “is-a” relation. The formal “is-a” relation is defined through subclass relations, whereas in informal systems it is expressed in natural language (Lassila and McGuinness 2001).

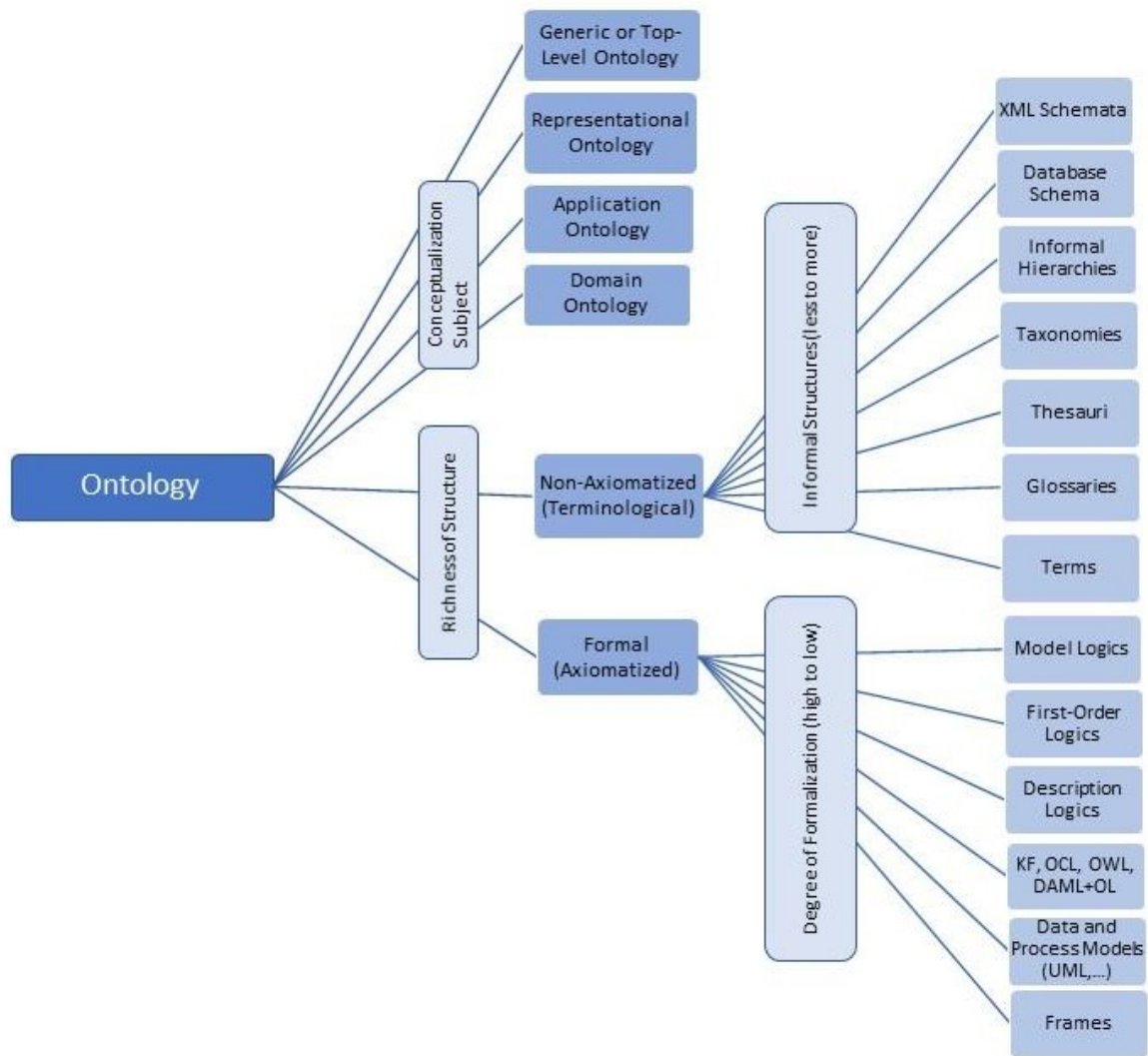


Figure 9: Ontology classification by degree of formalization (adapted from Gromann 2015:24)

The main scope of ontologies is to facilitate human-machine, inter-machine and inter-human communication (Guarino et al. 2009: 14). The communication scenario is represented by the semiotic triangle (See: 2.1. Terminologies (Odgen & Richards 1923)). In Figure 10, the semiotic triangle is extended to include the notion of context, which significantly influences communication. Within this representation, context connects the sign to the thing; this relation i.e. the arrow represents the act of reference, although it is always mediated by a mental concept. Context therefore determines the invoked concept, which refers to the object or phenomenon in reality. However, these relationships between sign, concept and thing remain ambiguous, since signs can be inappropriately associated to the wrong concepts.

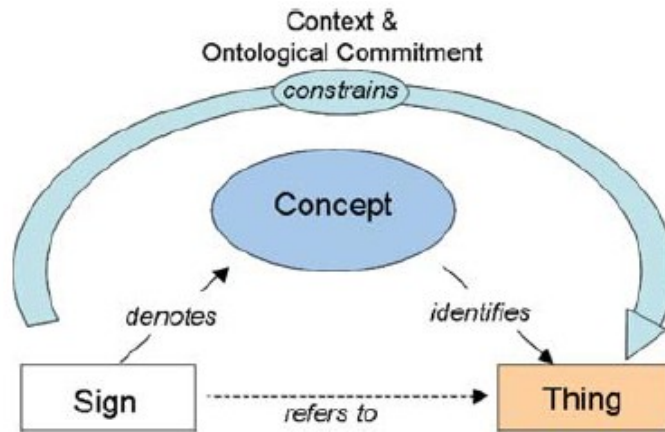


Figure 10: Semiotic triangle under an ontological perspective (Guarino et al. 2009: 16)

The misinterpretation of communication is heightened, when the ontological commitment is not *shared* among agents, meaning that no mutual agreement exists (Guarino et al. 2009: 14). As a results, different agents may infer divergent semantic interpretations and intentions from the same message. Conversely, if the commitment is shared to a common ontology, the inferred conclusions are restricted because the potential correspondences between signs, concepts and things become constrained (Guarino et al. 2009: 15f.).

In this regard, Drewer et al. (2017) describe the principal challenge in ontology building as the structuring and the naming of classes and instances. Synonymy and ambiguity hinder efficient knowledge exchange, because of a contextual misunderstanding. Furthermore, linking existing ontologies with each other represents another additional challenge, as different ontologies might employ different designations for the same concepts (synonymy) and they might also show structural differences, especially by granularity of the structure (Drewer et al. 2017: 23). These difficulties become even more pronounced in multilingual contexts, where ontologies described in different natural languages must be aligned. In such cases, machines are often unable to read and to understand the semantic and cultural embedded within different linguistic systems.

## 2.3 Semantic Web and Ontology Languages

The term Semantic Web (SW) was coined by Tim Berners-Lee in 2001 (Gromann 2015) and refers to the project aimed at enabling the documents ‘sharing and interconnection of each kind of format in the World Wide Web (WWW). These documents are described using formal languages and, in fact, represent a formal modelled knowledge. SW enables access to these

documents and supports the development of ontologies independent from natural languages, , through explicitly machine-readable semantics.

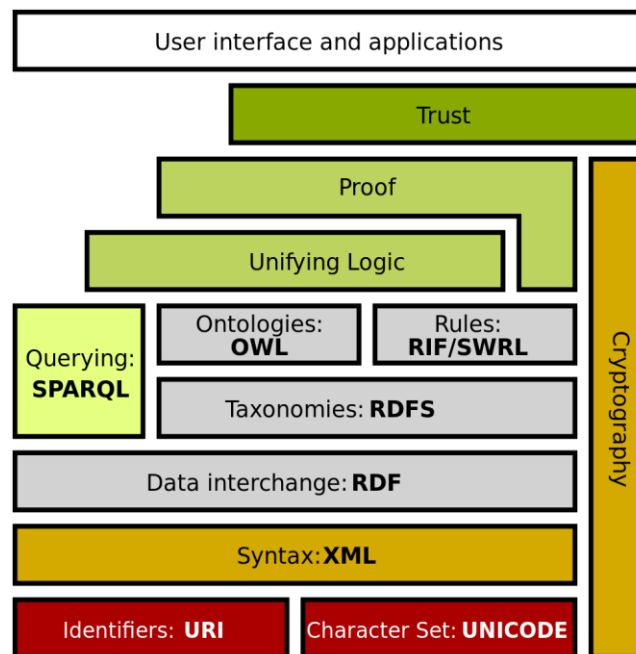


Figure 11: Semantic Web structure (Semantic Web Stack 2014)

The World Wide Web Consortium (W3C) is the standardisation organisation responsible for the development of Semantic Web technologies (Gromann 2015). W3C develops standards, guidelines and protocols to ensure the effective functioning of WWW, based on the principles, Web for All and Web on Everything, while aspiring to create a Web for Rich Interaction, Web of Data and Services and Web of Trust (W3C Mission). The primary objective of the Semantic Web, as defined by W3C, is enabling a better and useful exploitation of computer systems and “develop systems that can support trusted interactions over the network” (W3C Standards and Requirements). In fact, Semantic Web technologies allow the store of data on the Web, the creation of vocabularies and the formal axiomatization of rules governing stored data.

W3C (2004) describes the WWW as “the information space in which the items of interest, referred to as resources, are identified by global identifiers called Uniform Resource Identifiers (URI)”. In Figure 11 is illustrated the Semantic Web structure. At the core of the Semantic Web are Uniform Resource Identifiers (URIs), namely strings, that identify specific resources and permits their intercommunication within the WWW. URIs are therefore used to distinguish resources from one another (Hitzler et al. 2009). They refer to Uniform Resource Locators (URLs), i.e. to eXtensible Markup Language (XML) namespaces. XML provides the

syntactic framework for Semantic Web technologies (Gromann 2015: 22). While URIs rely on ASCII characters and are therefore limited in character range, Internationalized Resource Identifiers (IRIs) employ Unicode strings, which means that IRIs permit the use of a much wider range of characters, including non-Latin alphabet, such as Chinese, Cyrillic, Korean and Japanese (W3C 2014a).

The Resource Description Framework (RDF) represents information through directed graph composed of a set of nodes and edges, forming an RDF graph. RDF was developed to describe relationships between objects (Hitzler et al. 2009) “in a semantically structured way” (Gromann 2015: 22), thereby enabling machine-interpretable statements. Therefore, RDF is used to describe Web resources and constitutes the foundational building block of formal information representation, although it is itself not itself a fully formal representation language. RDF is used to make statements in the form of subject-predicate-object, referred to as triples. These semantic triples enable the codification of statements and the representation of knowledge in a machine-readable form. RDF graphs are composed of edges connecting three principal elements: a subject (IRI or blank node), a predicate (IRI) and an object (IRI, literal or blank node), although not every edge necessarily correspond to a valid RDF triple (Hitzler et al. 2009).

Resource Description Framework Schema (RDF Schema or RDFS) enlarge the RDF vocabulary, through a set of classes and properties. It allows the formulation of logic statements in RDF and supports the organization the knowledge in categories and subcategories. As results, they are used as formalism for lightweight ontologies (Hitzler et al. 2009).

The Web Ontology Language (OWL) is a standard language for modelling ontologies by W3C. Based on RDF, OWL describes the complex relationships between concepts, but it has the advantage to be easily readable by humans. As RDFS, OWL exploits categories and subcategories as core structural elements; however, it specifies much more complex and sophisticated semantic relations (Hitzler et al. 2009). In addition, OWL allows the formulation of expressions like predicate logic in a way that software can process semantically the meaning of the statement (Drewer et al. 2017).

The SPARQL Protocol and RDF Query Language (SPARQL) is the standard query language for RDF and it enables queries to be submitted in a form of simple graphs (Hitzler et al. 2009). It supports reasoning processes, that evaluate the stored knowledge by retrieving and

manipulating formal knowledge in RDF-base formal representations (Drewer et al. 2017). Thanks to this reasoning mechanism, systems can draw conclusions, so-called inferences, from existing knowledge.

Semantic Web technologies, in particular RDF, OWL, SPARQL, allow the linking of data, the querying of information, and the derivation of inferences. Linked Data refers to the network of interconnected dataset, created through the explicit description of relationships among data enties (W3C Standards and Requirements). Inference is enabled thanks to vocabularies, which define concepts and their relationships within a specific domain and is achieved through reasoning processes. Although no clear demarcation exists between vocabularies and ontologies in W3C Vocabularies, the distinction is generally made upon the degree of formality (Gromann 2015): ontologies are formal resources, whereas RDF vocabularies are comparatively informal.

## **2.4 Ontology Learning**

Gómez-Pérez & Manzano-Macho (2004) define the ontology learning (OL) as the set of methods and techniques employed for developing an ontology. Ontology learning process aims to extract ontologies from natural language (NL) texts semi-automatically with human intervention. Ontology learning (OL) is a multidisciplinary science, integrating linguistics, natural language processing (NLP), machine learning and knowledge representation (Mädche & Staab 2001).

Mädche & Staab (2004) identify three different possible sources for ontology learning: structured sources, semi-structured sources and unstructured data. Structured sources include existing ontologies, databases, term bases, lexicons etc., e.g. WordNet. Semi-structured sources comprise documents with predefined structure, such as XML texts. Unstructured sources consist of natural language texts.

When ontology learning is performed on unstructured data, different approaches are required in analysing NL text: linguistic techniques (e.g. linguistic annotation) is integrated with statistical methods and machine learning algorithms. In such cases, computational lexicons may serve as initial structural framework, that is enriched with new concepts and relations, obtained by the ontology learning process. Despite significant advances in the field, there is currently no universally accepted methodology or fully automatic tool for ontology learning

(Gómez-Pérez & Manzano-Macho 2004: 207). Nevertheless, most ontology learning architecture follow the general structure illustrated in Figure 12.

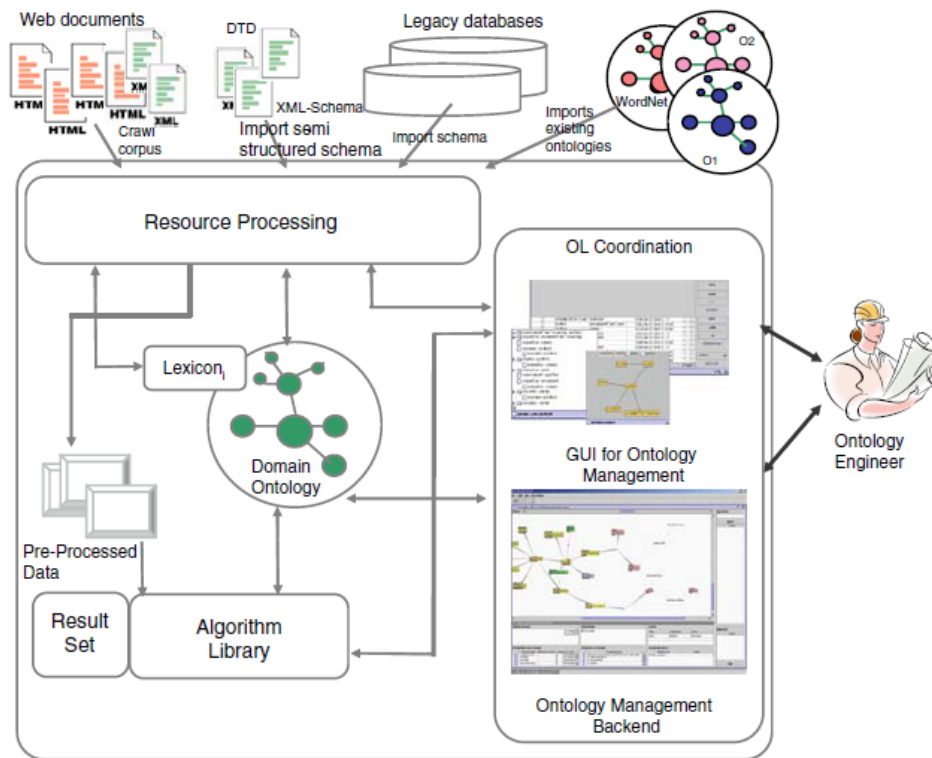


Figure 12: Ontology learning architecture (Cimiano et al. 2009: 247)

According to Cimiano et al. (2009) ontology learning architecture consists of four components. The first, ontology management, provides the interface between the ontology and the learning algorithms, enabling the addition of new concepts, relations or axioms to the ontology model. Furthermore, this component allows the inclusion of existing ontologies, their browsing, validation, modification, versioning and evolution (Mädche & Staab 2004).

The second component, coordination, manages the interaction between resource processing and algorithm library. It is responsible for selecting relevant data, applying transformation and preprocessing techniques, and activating specific extraction mechanisms (Mädche & Staab 2004).

The third component, resource processing, analyses input data, in order to generate a pre-processed set of data suitable for the algorithm library. A relevant sub-component of resource processing is the natural language processing system (Mädche & Staab 2004). The specific preprocessing strategy depends on the type of source data. For example, Hearst (1992) transformed semi-structured documents, such as dictionaries, into predefined relational

structures. In contrast, unstructured texts require to be processed through NLP system. One widely used NLP processors for English language is GATE (Cunningham et al. 1999). After linguistic analysis, the processed are transformed into algorithm-specific relational representations (Mädche & Staab 2004).

The fourth component, the algorithm library, contains the applied algorithms for the OL process. These algorithms are designed to extract, update and maintain the various elements of the ontology model (Mädche & Staab 2004). Since different algorithms may produce overlapping or complementary outputs, their results must be integrated, standardised, and consolidated.

Ontology learning aims to reconstruct a domain model from a textual input by exploiting the formal structures embedded within the data (Asim et al. 2008). The ontology layer cake illustrated in Figure 13 represents the different tasks or steps required to realize an ontology from unstructured text.

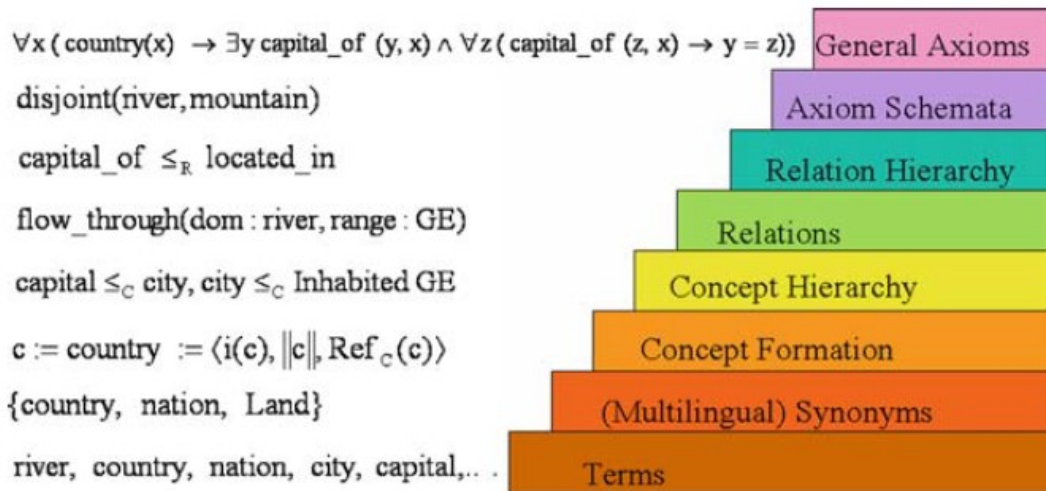


Figure 13: Ontology learning layer cake structure (Cimiano et al. 2009: 250)

In this model, each layer provides the input for the subsequent layer. The lowest level concerns the lexical level of ontology learning, i.e. the identification of terms and synonyms. These lexical units are subsequently used to generate concepts. Higher layers involve the acquisition of concept hierarchies, semantic relations and relation hierarchies. The uppermost levels, dedicated to the learning of axiom schemata and general axioms, represent the most challenging stage of ontology learning. Since the number of possible rules and axioms are potentially unlimited, complexity increases considerably. However, the adoption of a specific ontology

language restricts the set of admissible axioms and thus constrains the learning process (Cimiano et al. 2009).

### 2.4.1. Ontology Learning Techniques

In Asim et al. (2018) three different classes of ontology learning techniques are distinguished: statistical techniques, linguistic techniques and logical techniques. Statistical techniques rely on data mining, machine learning and information retrieval methods, used for extracting domain-specific terms, concepts and associations. Linguistic techniques (see 2.4.1.1. Linguistic techniques) are based on natural language processing methods, whereas logical techniques employ formal representations and Inductive Logic Programming (ILP – see 2.4.1.3. Logic technique: ILP). Figure 14 schematizes the ontology learning process and classifies these techniques according to colour coding: blue is used for detecting statistical techniques, red for linguistic and purple for logical techniques.

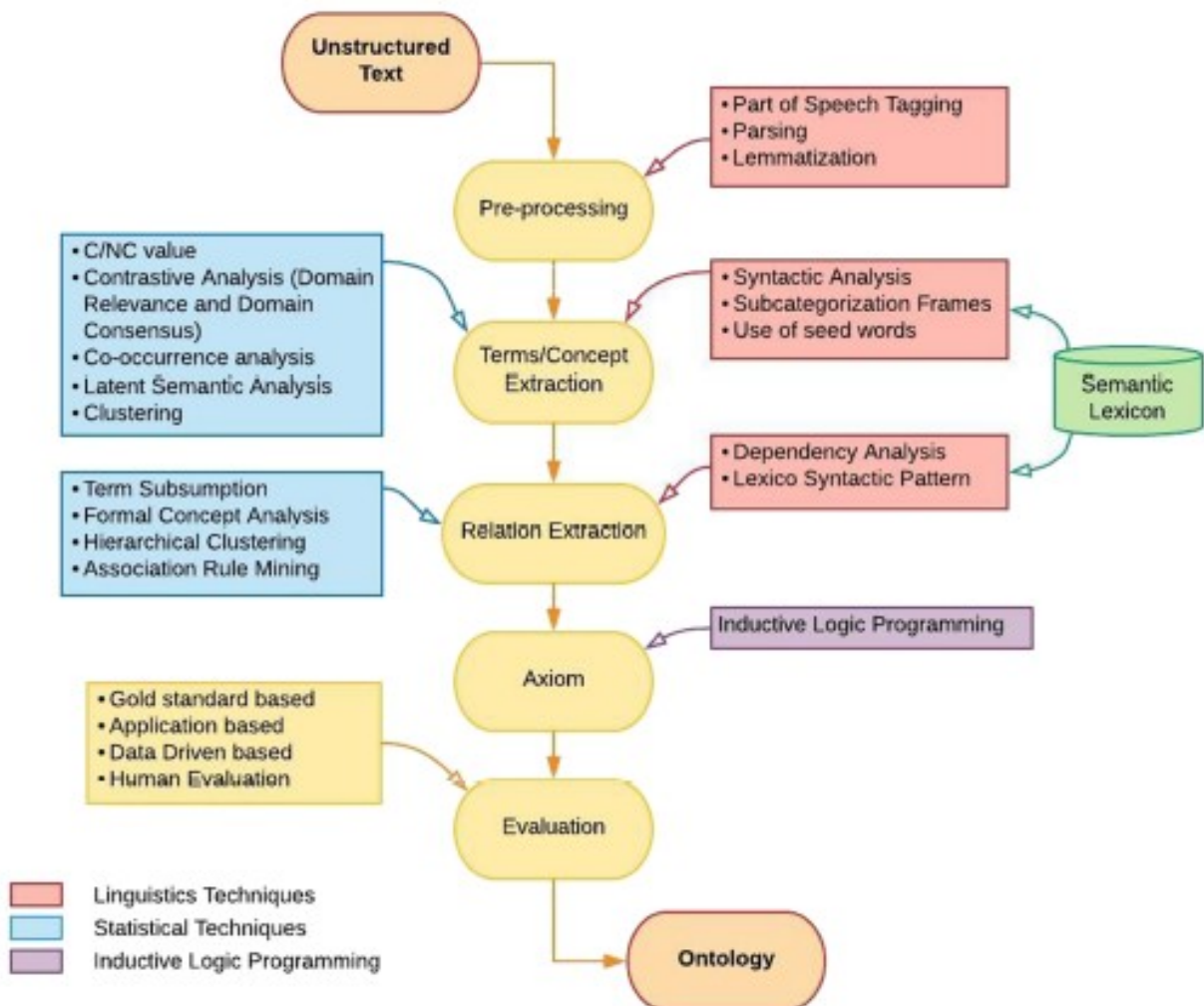


Figure 14: Ontology learning techniques (Asim et al. 2018: 4)

#### *2.4.1.1. Linguistic Techniques*

Linguistic techniques are employed both for texts preprocessing and for extraction of terms, concepts and relations. Jiang & Tan (2010) state the importance of pre-processing, so that the ontology learning task achieves a higher accuracy. Preprocessing consists of three stages: part-of-speech tagging, sentence parsing and lemmatization.

Part-of-speech tagging assigns grammatical categories, such as nouns, verbs, and adjectives, to lexical items within a corpus (Asim et al. 2018). Syntactic parsing analyses grammatical interdependencies among words and represents them through parsing trees that capture relationships such as subject-object structures (Asim et al. 2018). In case of unstructured data, knowledge acquisition is still problematic and generally involves three stages: grammatical analysis, formalisation and integration with lexical resources.

The sentence parsing aims to identify structural dependencies within sentences through tasks such as part-of-speech tagging, named entity recognition, chunking and word-sense disambiguation (Hitzler et al. 2009: 312). However, parsing remains challenging because algorithms are highly language-dependent, and ambiguities, particularly those involving pronouns, may lead to incorrect interpretations. In the second stage, parse trees are converted into logical statements, a process referred to formalisation. Finally, the resulting structures are integrated with lexical resources such as thesauri and WordNet (Hitzler et al. 2009).

Lemmatization identifies the canonical forms of inflected lexical items. This process reduces extension of data and improves the retrieval of relevant terms, thereby facilitating subsequent ontology learning tasks (Drymonas 2009).

Knowledge extraction, including the extraction of terms, concepts and relations, can be performed through syntactic structure analysis, subcategorization frames, and seed-words approaches. Syntactic structure analysis relies on dependency analysis and lexico-syntactic patterns (Ciaramita et al. 2005; Hearst 1998; Sombatsrisomboon 2003; Snow et al. 2005; Sen 2015; Buitelaar et al. 2005; Kaushik & Chatterjee 2018; Ismail et al. 2015; Ismail et al. 2017; Panchenko et al. 2016; Atapattu et al. 2017). Initially, corpora are annotated with part-of-speech tags, in order to extract syntactic structures as noun phrases and verb phrases.

Subcategorization frames identify the lexical environments in which a word typically occurs and thereby facilitate the identification of related semantic classes (Asim et al. 2018: 6).

For example, in the sentence “Anna writes a letter”, the subcategorization frames of “write” are “Anna” and “a letter”, which belong to the classes “Person” and “written communication medium”. As a result, the frame “write” will be associated with these two semantic categories (Agustini et al. 2001; Gamallo et al. 2003). Furthermore, Gamallo et al. (2003) state when combined with clustering techniques, subcategorisation frames can also support concept extraction.

Seed-word approaches rely on domain-specific lexical items from which algorithms identify semantically related terms and concepts (Hwang 1999). Their introduction of seed-words has significantly improved concept and term extraction processes in the OL process (Asim et al. 2018).

Relation extraction is typically performed through dependency analysis or lexico-syntactic patterns. Dependency analysis exploits syntactic dependencies represented in parsing trees to identify semantic relationships among concepts (Kang et al. 2015; Ciaramita 2005). Dependency analysis exploits dependency path information represented in parse trees to identify relationships among concepts (Ciaramita 2005). Sordo et al. (2015), by contrast, exploit the lexico-syntactic patterns that can be more suitable for extracting taxonomic, non-taxonomic and is-a relations. This pattern extraction, although, requires massive efforts, so that Snow et al. (2005) apply machine learning algorithm for automatic patterning extraction. Logistic regression is applied to learned hypernym pairs. Thus, their dependency relationships are exploited, in order to extract new relationships from unknown data.

A notable example is the use of lexico-syntactic patterns by Hearst (1992), which exploit recurring linguistic structures to identify semantic relations. It employs a heuristic method: it exploits the regular expressions that reoccur in various texts and, as the text is scanned, the sentence similar to the semantic structure of the pattern can be detected. For instance, the pattern “x is\_a y” can be utilized to learn the hierarchical relation that y is a superordinate concept of x. Applied to the sentence “a bird is an animal”, this pattern enables the extraction of hierarchical taxonomic relation. In such a manner semantic relations are highlighted, resulting in a taxonomy (Mädche & Staab 2004).

Semantic lexicons appear in green in Figure 14, since they can be applied both in knowledge extraction and in relation extraction processes. WordNet is one of the most applied

semantic lexicons, which are used as knowledge resources, since they provide predefined concepts and relations (Asim et al. 2018).

#### 2.4.1.2. Statistical Techniques

Statistical techniques are primarily employed for term and concept extraction, as well as for identification of taxonomic relations. Unlike linguistic approaches, these methods focus on the frequencies and probabilities of lexical items, rather than on the semantic interpretation of corpora. Term and concept extraction can be conducted through different techniques: term frequency, C/NC value, contrastive analysis, co-occurrence analysis, and Latent Semantic Analysis (LSA) algorithms and clustering.

One of the central challenges in ontology learning is the identification of relevant terms and concepts. The development of ontologies often relies on the reuse of existing algorithms and their adaptation through new definitions and implementations.

Traditional frequency-based methods assume that the more frequently a term or concept occurs in a corpus  $D$ , the more relevant it is assumed to be to the target domain. However, this assumption has been reassessed, and the integration of information retrieval methods, such as term weighting, has demonstrated superior performance compared to simple frequency counts.

One of the most widely adopted weighting methods is Term Frequency–Inverse Document Frequency (TF–IDF) (Mädche & Staab 2004). Let:

- $L$  denote the set of lexical entries;
- $D$  denote the document corpus;
- $|D|$  denotes the total number of documents in the corpus;
- $l \in L$  be a lexical entry;
- $d \in D$  be a document;
- the term frequency  $tf_{l,d}$ , representing the frequency of occurrence of lexical entry  $l \in L$  in a document  $d \in D$ ;

- the document frequency  $df_l$ , representing the number of documents in the corpus  $D$  in which lexical entry  $l$  occurs;
- the corpus frequency  $cf_l$ , representing the total number of occurrences of  $l$  in the overall corpus  $D$ ;
- $\mathbf{R}^+$  denotes the set of positive real numbers.

The following relations hold:  $df_l \leq cf_l$ , and  $\sum_{d \in D} tf_{l,d} = cf_l$ .

The TF-IDF value of a lexical entry  $l$  in a document  $d$  is calculated as:

$$tfidf_{l,d} = tf_{l,d} * \log \left( \frac{|D|}{df_l} \right).$$

This weighting scheme reduces the relevance of terms that occur either too frequently or too infrequently within a corpus, thereby highlighting terms that are particularly characteristic of specific documents (Mädche & Staab 2004: 180). To estimate the overall relevance of a lexical entry across the entire corpus rather than within a single document, the TF-IDF values are aggregated over all documents. The corpus-level TF-IDF value is defined as:

$$tfidf_l = \sum_{d \in D} tfidf_{(l,d)}$$

where  $tfidf_l \in \mathbf{R}^+$ .

The standard TF-IDF algorithm does not account for stop word lists. Therefore, a positive threshold  $k \in \mathbf{R}^+$  may be defined such that only lexical entries satisfying  $tfidf_l > k$  are retained for further analysis and potential inclusion in the ontology's concept hierarchy and lexical inventory (Mädche & Staab 2004: 180).

Another widely used technique for terminology extraction is the C/NC-value method, which combines the C-value and NC-value measures for the extraction of multi-word terms (Frantzi et al. 2000). The C-value identifies candidate terms that are statistically significant within a corpus, whereas the NC-value extends this measure by incorporating contextual information. (Asim et al. 2018: 7). Once the C value is identified, context words are extracted as a list. Specifically, context words located immediately to the left and right of a candidate

term are extracted and assigned weights (Asim et al. 2018: 7). These contextual weights are subsequently combined with the C-value to obtain the NC-value. Frantzi et al. (2000) demonstrated that the C-value improves the precision of term extraction, while the NC-value further enhances the identification of valid domain-specific terms.

Contrastive analysis is another statistical approach used to filter out terms extracted but not specific to a target domain (Navigli & Velardi 2002). Navigli et al. (2003) distinguish two different processes of contrastive analysis: domain relevance, which computes the “specificity” of a term within the target domain, and domain consensus, which identifies terms that occur consistently across a substantial number of documents within that domain. This approach requires two corpora: a target-domain corpus and a contrastive corpus representing a non-relevant domain (Asim et al. 2018: 7).

Co-occurrence analysis is employed to identify concepts through the detection of implicit associations between terms and concepts. Similarly, the Latent Semantic Analysis (LSA) is based on the assumption that terms that frequently occur together are likely to share semantic properties (Asim et al. 2018). Another important statistical technique is clustering, “an unsupervised learning approach” (Asim et al. 2018: 8), that groups objects according to similarity measures (Berkhin 2006). Clustering techniques are frequently applied to concept extraction and taxonomy construction.

Relation extraction can be made through different statistical approaches: term subsumption, Formal Concept Analysis (FCA), hierarchical clustering and association rule mining (ARM). Term subsumption identifies hierarchical relations between terms by analysing their distribution across documents. This algorithm compares two terms, one is more general than the other, by identifying them within corpora. If all occurrences of term  $x$  are found within documents that also contain term  $y$ , then  $y$  is considered more general than  $x$ . This technique is used for non-specialist terms and to extract general relationships between terms (Njike-Fotzo & Gallinari 2004).

Formal concept analysis (FCA) is used to build concept hierarchies by analysing object-attribute matrices. FCA analyses the object attribute matrix and finds the clusters related to those attributes and objects, and subsequently it generates a hierarchical relationship among concepts and attributes.

Hierarchical clustering is one of the most frequently applied techniques for determining taxonomic relations between data. Terms are grouped into clusters according to similarity measure, enabling the extraction of concepts and the construction of hierarchical structures. Two principal approaches can be distinguished: agglomerative clustering or bottom-up approach and divisive clustering or top-down approach (also partitional approach in Mädche & Staab 2004). In agglomerative clustering, each element initially forms an “individual cluster” and similar clusters are iteratively merged (Zepeda-Mendoza & Resendis-Antonio 2013). In divisive clustering, all elements initially belong to a single cluster, which is progressively subdivided into smaller clusters, creating a hierarchy (Dhillon et al. 2003). Mädche & Staab (2004: 181) additionally discuss a third approach, the conceptual clustering, which constructs networks of terms by analysing overlaps of representing terms between two represented terms.

Finally, Association Rule Mining (ARM) is a data mining technique designed to identify hidden relationships and association patterns among several elements in a dataset. ARM is primarily used for the extraction of non-taxonomic relations and for discovering rules that predict the co-occurrence of concepts within a database.

#### 2.4.1.3. Logic technique: Inductive Logic Programming

Inductive Logic Programming (ILP) is employed at the highest level of the ontology layer cake, where the objective is the generation of axioms based on existing background knowledge. ILP combines learning with formal logic to infer new rules and relationships from observed data.

The applications of ILP in ontology learning are various. For example, it has been used for ontology population from web resources through the integration of lexical databases such as WordNet and domain-independent linguistic patterns (Lima et al. 2013). Furthermore, ILP has been applied to the extraction of domain-specific terminology for the construction of topic ontologies from textual corpora (Fortuna et al. 2008). By enabling the automatic generation of axioms and formal rules, ILP contributes significantly to the enrichment and formalisation of ontological knowledge structures.

## 2.5 FRED: Machine Reading for Semantic Web

FRED (Gangemi et al. 2013b) is a knowledge extraction (KE) system, often referred to as a machine reader, which transforms natural language (NL) texts into formal knowledge representations using Semantic Web technologies. Specifically, FRED generates RDF/OWL

output graphs by combining natural language processing (NLP) techniques with external semantic resources, including existing Semantic Web ontologies such as VerbNet, FrameNet, WordNet, DBpedia, and foundational ontologies. Furthermore, FRED incorporates RDF encodings of syntactic annotations through the use of the Extreme Annotational RDF Markup Language (EARMARK) (Peroni & Vitali 2009) and the NLP Interchange Format (NIF) vocabularies (Hellmann et al. 2013).

The system exemplifies the interaction between semantic technologies and NLP by exploiting the relationship between formal knowledge structures, represented by computational ontologies, and linguistic knowledge. In this context, linguistic information is enriched through access to formal semantic representations, while formal knowledge bases benefit from the integration of linguistic data (Gangemi 2013).

### **2.5.1 Software Structure**

According to Presutti et al. (2012: 126f.), FRED consists of four main components represented in Figure 15. The first component is Boxer (Bos 2008), a semantic parser that analyses NL text, including frame detection. Boxer employs Combinatory Categorical Grammar (CCG) (Steedman 2000) and Discourse Representation Theory (DRT) (Kamp 1981) to generate formal semantic representations based on event semantics. The resulting DRT-based document representation is subsequently transformed into OWL/RDF ontologies through a set of heuristics rules implemented by FRED (Gangemi et al. 2013a). One notable advantage of Boxer usage is that it dispenses the need of ad-hoc-training procedures typically required by machine-learning-based ontology

The second is communication component, which realizes a lightweight Hypertext Transfer Protocol (HTTP) server based on Representation State Transfer (REST) architecture (Fielding 2000). Natural language text is submitted as input and, following processing by FRED, is returned as an RDF/OWL ontology.

The third is the refactoring component, which converts the output produced by Boxer into a more structured and readable format suitable for subsequent processing by the re-engineering component.

The last is the reengineering component, which implements translation and heuristic rules to transform the DRT representation generated by Boxer into an OWL ontology. Through this process, formal semantic structures are mapped onto Semantic Web representations.

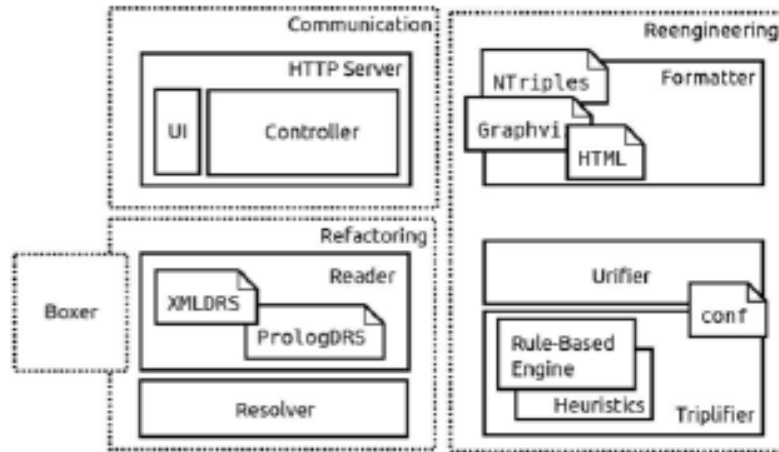


Figure 15: FRED Software Architecture (Gangemi et al. 2013a)

## 2.5.2 Tasks and Functional Properties

Gangemi (2013) compared several knowledge extraction tools, available at the time and reported that FRED achieved a precision of 0.73, a recall of 0.93, and an overall accuracy of 0.87. These results substantially outperform those of other event extraction systems evaluated in the study (Gangemi et al. 2013b).

In this context, precision refers to the proportion of correctly extracted elements, whereas recall denotes the proportion of elements successfully extracted from the complete set of elements present in the source text.

Gangemi (2013) identifies twelve fundamental tasks that a machine reading tool should support, establishing correspondences between NLP operation and Semantic Web representations:

1. topic extraction (recognition of domain-specific topics, e.g. individuals belonging to the range of the property `dc:subject`)
2. named entity recognition (individual induction)
3. named entity resolution (identity resolution of individuals)
4. named entity coreference (coreference of individuals)

5. terminology extraction (induction of constants pertinent to a domain, typically for classes or properties)
  - a. class induction
  - b. property induction
6. sense tagging (or class membership induction)
7. sense disambiguation (also identity resolution for classes)
8. taxonomy induction (also subclass relation induction)
9. (non-taxonomic, non-role, binary) relation extraction (property assertion – fact–induction)
10. semantic role labelling (also property induction for events and n-ary relations)
11. event detection (also n-ary relationship induction)
12. frame detection (also n-ary relation –type– induction)

Figure 16 summarizes which processes are required and applied in the different tools to perform linguistic analysis and to obtain formal semantics (KE applied in SW) and shows the final results of the comparison conducted in Gangemi (2013):

| Tool           | Topics | NER | NE-RS | TE | TE-RS | Senses | Tax | Rel | Roles | Events | Frames |
|----------------|--------|-----|-------|----|-------|--------|-----|-----|-------|--------|--------|
| AIDA           | –      | +   | +     | –  | –     | +      | –   | –   | –     | –      | –      |
| Alchemy        | +      | +   | –     | +  | –     | +      | –   | +   | –     | –      | –      |
| Apache Stanbol | –      | +   | +     | –  | –     | +      | –   | –   | –     | –      | –      |
| CiceroLite     | –      | +   | +     | +  | +     | +      | –   | +   | +     | +      | +      |
| DB Spotlight   | –      | +   | +     | –  | –     | +      | –   | –   | –     | –      | –      |
| FOX            | +      | +   | +     | +  | +     | +      | –   | –   | –     | –      | –      |
| FRED           | –      | +   | +     | +  | +     | +      | +   | +   | +     | +      | +      |
| NERD           | –      | +   | +     | –  | –     | +      | –   | –   | –     | –      | –      |
| Open Calais    | +      | +   | –     | –  | –     | +      | –   | –   | –     | +      | –      |
| PoolParty KD   | +      | –   | –     | –  | –     | –      | –   | –   | –     | –      | –      |
| ReVerb         | –      | –   | –     | –  | –     | –      | –   | +   | –     | –      | –      |
| Semioseach     | –      | –   | +     | –  | +     | –      | –   | –   | –     | –      | –      |
| Wikimeta       | –      | +   | –     | +  | +     | +      | –   | –   | –     | –      | –      |
| Zemanta        | –      | +   | –     | –  | –     | –      | –   | –   | –     | –      | –      |

Figure 16: Comparison of Machine Reading Tools (Gangemi 2013: 360)

Unlike most contemporary machine reading systems, FRED does not rely primarily on machine learning techniques. Instead, its theoretical foundation is based on the logical interpretation of natural language derived by Discourse Representation Theory (DRT) (Presutti et al. 2012). Furthermore, FRED is unique in its ability to generate RDF graphs and creates event frames through the integration of external semantic resources, such as VerbNet and FrameNet (Gangemi 2013).

Gangemi et al. (2013b) characterize FRED as an event-centric system, where event is interpreted as an elusive identity, the extension of event's participants. That means events are considered the central entities around which semantic representations are organized. Syntactically, events are typically expressed by verbs and are interpreted as structure connecting participant, roles and contextual information.

FRED implements five principal event-related functions:

1. Event identity: events are expressed by verbs, propositions, common nouns and named entities, such as proper nouns.
2. Event classification: Identified events are linked to existing Semantic Web ontologies through Linked Data resources.
3. Event unity: event boundaries (labelling) and semantic roles (frame detection) are defined through frame detection and mapped to a shared event ontology.
4. Event modifiers: logical negation, basic modalities, and adverbial qualities are extracted.
5. Event relations: temporal and logical relationships among events are extracted through the analysis of verbs and propositions.

### **2.5.3 Results of FRED**

The output of FRED can be either a graphical representation or textual formats. Supported serialisation formats include Terse RDF Triple Language (Turtle), RDF/XML, RDF/JSON, N3, NT, and Directed Acyclic Graphs (DAGs). The output consists of RDF triples, containing both Semantic Web fixed properties, such as `rdf:type`, `rdfs:subClassOf`, `owl:sameAs`, `dul:associateWith`, `owl:equivalentTo` and FRED-specific properties representing extracted semantic roles and events (Gangemi et al. 2013a).

Indeed, FRED graphs include both EARMARK and NIF annotations that establish explicit links between textual fragments and their matching graph elements. These annotations facilitate interoperability among NLP tools and support integration of heterogeneous semantic resources (Gangemi et al. 2017). Figure 17 lists the prefixes used in FRED either in RDF/OWL graphs or in Turtle.

|                                          |            |                                                                                                                                             |
|------------------------------------------|------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Default local (customisable) namespace   | fred:      | <a href="http://www.ontologydesignpatterns.org/ont/fred/domain.owl#">http://www.ontologydesignpatterns.org/ont/fred/domain.owl#</a>         |
| VerbNet thematic roles                   | vn.role:   | <a href="http://www.ontologydesignpatterns.org/ont/vn/abox/role/">http://www.ontologydesignpatterns.org/ont/vn/abox/role/</a>               |
| VerbNet verb classes                     | vn.data:   | <a href="http://www.ontologydesignpatterns.org/ont/vn/data/">http://www.ontologydesignpatterns.org/ont/vn/data/</a>                         |
| FrameNet frame vocabulary                | ff:        | <a href="http://www.ontologydesignpatterns.org/ont/framenet/abox/frame/">http://www.ontologydesignpatterns.org/ont/framenet/abox/frame/</a> |
| FrameNet frame element vocabulary        | fe:        | <a href="http://www.ontologydesignpatterns.org/ont/framenet/abox/fe/">http://www.ontologydesignpatterns.org/ont/framenet/abox/fe/</a>       |
| DOLCE+DnS Ultra Light                    | dul:       | <a href="http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#">http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#</a>                 |
| WordNet                                  | wn30:      | <a href="http://www.w3.org/2006/03/wn/wn30/instances/">http://www.w3.org/2006/03/wn/wn30/instances/</a>                                     |
| a vocabulary for Boxer primitive classes | boxer:     | <a href="http://ontologydesignpatterns.org/ont/boxer/boxer.owl#">http://ontologydesignpatterns.org/ont/boxer/boxer.owl#</a>                 |
| a vocabulary for some box types          | boxing:    | <a href="http://ontologydesignpatterns.org/ont/boxer/boxing.owl#">http://ontologydesignpatterns.org/ont/boxer/boxing.owl#</a>               |
| DBpedia resources                        | dbpedia:   | <a href="http://dbpedia.org/resource/">http://dbpedia.org/resource/</a>                                                                     |
| schema.org vocabularies                  | schemaorg: | <a href="http://schema.org/">http://schema.org/</a>                                                                                         |

Figure 17: List of prefixes used in FRED diagrams and inline examples (Gangemi et al. 2017: 4)

The RDF/OWL graphs generated by FRED are grounded in several theoretical frameworks, including Combinatory Categorical Grammar (CCG), Discourse Representation Theory (DRT), Discourse Representation Structures (DRSs) (Seedman 2000), Linguistic Frames (Fillmore 1976) and Ontology Design Patterns (ODP) (Gangemi 2005). Since the generated graphs are generated from NL texts, they are considered largely “independent from any commitment on domain-specific or task-specific semantics” (Gangemi et al. 2017: 2). Consequently, FRED can function effectively as a software semantic mediator across different domains.

Through Boxer, NL input is analysed according to CCG principles and transformed into event-centred semantic representations, i.e. verb-centred. These representations are subsequently formalised using DRT, a formal semantic theory equivalent in expressive power to first-order logic (Presutti et al. 2012).

Discourse Representation Structures (DRSs) consist of two principal components: discourse referents, and conditions that specify the interpretation of discourse referents. Together, these elements provide a formal representation of sentence meaning (Gangemi et al. 2017). In the following, Figure 18 shows the FRED’s graphic representation output for the sentence “Valentina gave Aldo a book by Charlie Mingus”.

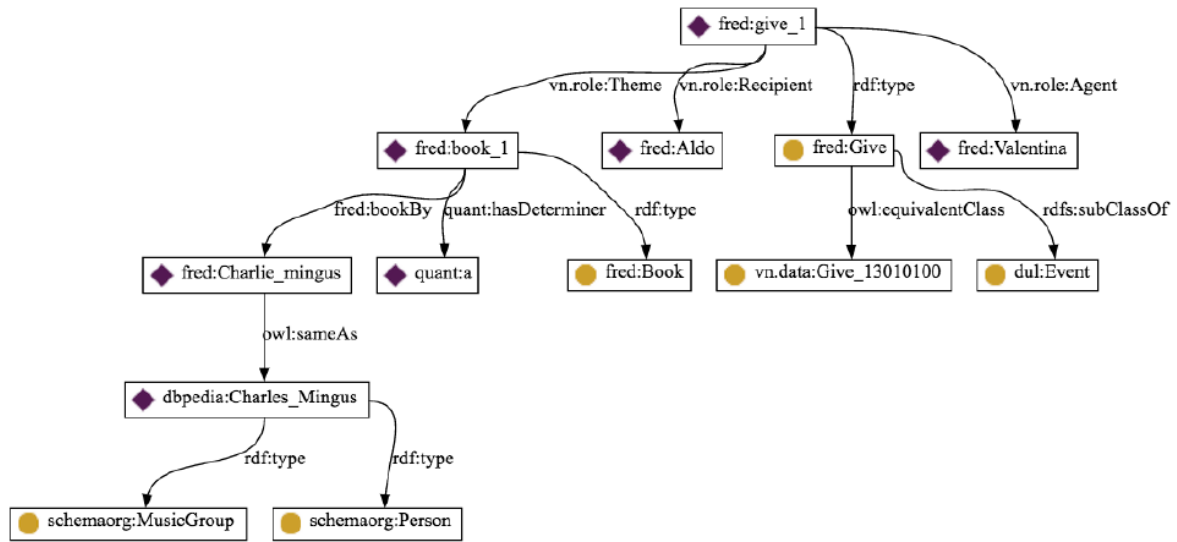


Figure 18: FRED's output (Gangemi et al. 2017: 3)

#### 2.5.4 Theoretical Foundations of FRED

FRED is constructed on basis of two main theoretical foundations: Frame Semantics and Ontology Design Patterns. Frame Semantics provides one of the theoretical foundations of FRED. According to this theory is understanding a lexical item requires access to a broader contextual and/or encyclopaedic knowledge associated with it.

Fillmore (1976: 27f.) argues that “the process of understanding a word requires us to call on our memories of experiences [...] through which we have learned the words in their labelling or describing functions”. Consequently, linguistic frames represent prototypical situations that provide contextual knowledge about events, objects, and participants (frame occurrences).

Presutti et al. (2012) describe how FRED operationalises Frame semantics by representing real-world knowledge thanks to semantic frames. These frames describe and characterize situations, objects, and events together with their associated frame elements, namely semantic roles. Frame detection therefore plays a crucial role in identifying complex syntactic and semantic relationships within NL texts. Frame detection therefore plays a crucial role in identifying complex semantic and syntactic relationships within texts and constitutes one of the most sophisticated aspects of ontology construction in FRED.

Janowicz et al. (2016) define pattern as “distinctive and repetitive invariants across observed data”. Ontology design patterns (ODPs) (Gangemi 2005) formalise such recurring

structures as reusable modelling solutions. ODPs support various Semantic Web applications, including semantic interoperability, knowledge extraction, query answering, and Linked Data publication and integration.

### 2.5.5 Translation from DRT to OWL

Although Boxer automatically generates formal semantic representations, these differ significantly from RDF/OWL ontologies at the syntactic, lexical, and semantic levels. Consequently, a dedicated set of translation rules is required to map DRT structures onto OWL representations (Presutti et al. 2012). Table 1 summarizes the main translation rules:

Table 1: Main Translation Rules from DRT to OWL (adapted from Presutti et al. 2012: 123)

| DRT Construct      | Boxer Syntax                                                | FOL Construct                         | OWL Construct                                   |
|--------------------|-------------------------------------------------------------|---------------------------------------|-------------------------------------------------|
| Predicate          | <code>pred(x)</code>                                        | Unary predicate $\phi$                | <code>rdf:type</code>                           |
| Relation           | <code>rel-name(x, y)</code>                                 | Binary relation                       | <code>owl:ObjectProperty</code>                 |
| Eq Relation        | <code>eq(x, y)</code>                                       | Equality predicate                    | <code>owl:sameAs</code>                         |
| Named Entity       | <code>named(&lt;var&gt;, &lt;name&gt;, &lt;type&gt;)</code> | Unary predicate $\phi$                | <code>owl:NamedIndividual</code>                |
| Discourse Referent | <code>&lt;var&gt;</code>                                    | Quantified variable                   | (generated)<br><code>owl:NamedIndividual</code> |
| DRS                | <code>&lt;drs&gt;</code> with event $E$                     | Proposition $P$ with predicate $\phi$ | RDF graph $G_p$ with class $E$                  |
| Negated DRS        | <code>not(&lt;drs&gt;)</code>                               | Negated proposition $\neg P$          | $G_p$ with<br><code>owl:disjointWith E</code>   |

The resulting FRED graphs are fundamentally frame- and event-centric. Their object properties correspond to recognised semantic roles and relationships. When FRED detects a semantic role that cannot be matched to an existing ontology, it automatically generates a new object property derived from the input text. Conversely, when individuals can be linked to existing Semantic Web resources, they are represented using established classes and properties from ontologies such as `dbpedia:` (Gangemi et al. 2017: 2).

Finally, FRED supports the analysis of 48 different languages. However, all input texts are first translated into English before semantic processing takes place. Consequently, generated RDF graphs always contain English labels. For non-English input texts, users must prepend the language identifier tag `<BING_LANG:lang>`, where `lang` corresponds to the appropriate ISO language code (Gangemi et al. 2017).

## 2.6 Ontologies and Terminologies in Comparison

To clarify the differences that exist between terminologies and ontologies, it is necessary to consider the three levels of ontological realism proposed by Smith et al. (2006): first, real world objects, processes, qualities, states, i.e. the objective reality; second, cognitive representation of reality; and third graphical representation of level 1 and 2. While the first level corresponds primarily to the domain of ontology, the second and third levels are concerned with the communication and representation of knowledge and are therefore closely related to terminology.

Thus, the main concern about the treated research field is the complications resulting from the contrast between the being itself and the cognitive, communicative and informational representation of the real-world (Budin 1996). This contrast, referred to by Bodenreider et al. (2004) as a *cleft*, arises from the relation between symbol and referent within the semiotic triangle (Smith et al. 2006), as this is based on thoughts and/or reference, subjectively indeed. Real-world objects are not directly linked to terms, nor do terms objectively correspond to real-world entities. Rather, concepts, *mental orienting elements*, are the prerequisite for every communication, functioning as mediating cognitive structures that connect terms and real-world objects. In this regard, Budin (1996: 22) writes that the epistemic structures establish the relationship between *onta* (gr. objects) and *epistemata* (gr. concept), thereby enabling communication by linking them to linguistic or to non-linguistic units.

But this can become an issue in terms of computer science. Human knowledge is inherently language-dependent, and natural languages continuously evolve as knowledge develops. Ontologies, however, aim to describe the world independently of natural language through formal ontology languages and axioms. Consequently, all conceptual relationships must be explicitly defined. Gromann (2015: 13) therefore argues that „[o]ntologies based on formal (logical) design principles are categorically separate from any such terminological or terminographic system”.

At the same time, effective communication between humans and machines necessarily occurs through natural language. As Drewer et al. (2017) observe, that the communication between human and machine has to take place in natural language, i.e. the machine needs to understand human unstructured data. Indeed, natural language expressions in ontologies are considered non-logical, non-abstract denominations of entities (Gromann 2015: 15).

Although both terminologies and ontologies employ signs as medium signs for organizing and representing knowledge, they study reality and domain-specific knowledge differently, searching for different results and different objects. Both disciplines root from Aristotelian logic and both aim at knowledge management as “shared conceptualization” (Gromann 2015: 61), but terminologies focus on how terms are used in specialized communication and which are the terminological differences within and across languages (second and third level of ontological realism), while ontologies describe formally a certain domain (first level of ontological realism).

Over recent decades, terminological resources have increasingly adopted characteristic associated with ontologies (Gromann 2015: 62f.). Despite this convergence, a fundamental distinction remains. Ontologies are designed to be interpreted by machines, whereas terminologies aim to define concepts and their interrelations in a manner that is comprehensible to human users (Drewer et al. 2017). Certainly, both terminology science and computer science will benefit from the synergy between ontology engineering and lexical semantics. Efforts to automatic conversion of terminologies to ontologies and vice versa, as well as attempts to develop integrated ontology-terminology models, have therefore become important research objectives in linguistics, computer science, and knowledge engineering.

The present study adopts a human-centred perspective and examines the relationship between terminologies and ontologies through a semiotic framework proposed by Gromann (2013, 2015). The semiotic approach facilitates a detailed analysis of signs, their representations and the modelling of linguistic patterns for specific scopes (Sowa 2000). A similar comparison has been conducted yet by Kless & Milton (2010), that compared thesauri and ontologies under a semiotic perspective. Bodenreider et al. (2004) and Klein & Smith (2010) evaluated the differences, but from an ontological point of view.

Semiotics investigates signs, their relations to real-world objects, their conceptualization, and their designation. As such, it provides an appropriate framework for analysing the representation, organisation, and use of signs in both terminologies and ontologies. Traditionally, semiotics is divided into three branches: syntax, which examines the representation of signs and their interrelations. Semantics, which studies meaning, reference and extension of concepts; and finally, pragmatics, which studies the use of signs in context and their practical effects. Gromann (2015: 107) translates these disciplines into practical

categories as form, meaning and usage of signs, i.e. scope and motivation, as exemplified in Table 2.

Table 2: Criteria for Semiotic Comparison of Terminologies and Ontologies (adapted from Gromann 2015)

|                        |                                                      |
|------------------------|------------------------------------------------------|
| 1) Syntax = Form       | a) mechanisms for grouping objects and relating them |
|                        | b) permissible content of those groupings            |
|                        | c) operations permissible on those groupings         |
| 2) Semantics = Meaning | a) mechanisms for expressing meaning                 |
|                        | b) mechanisms for interpreting meaning               |
|                        | c) relation to real-world objects                    |
| 3) Pragmatics = Use    | a) purpose of modeling intended meaning              |
|                        | b) development process                               |
|                        | c) application scenarios                             |

### 2.6.1 Syntax

The syntactic comparison between terminologies and ontologies is based on the standard recommendations of ISO and World Wide Web Consortium (W3C) are identified and compared. Specifically, the languages compared are TBX (TermBase eXchange) (LISA 2008) for terminologies and OWL (Web Ontology Language) for ontologies. Both standards exist in different variants; TBX comprises several dialects, whereas OWL provides different syntaxes. Therefore, by comparing their structural components, it is possible to identify those elements that match fully according to their definitions, those that share a partial correspondence and those with no correspondence.

Table 3 presents the results of comparison proposed by Gromann (2015). Elements that appear in both the terminology and ontology columns represent full correspondences. Elements enclosed in parentheses indicate partial correspondences, whereas elements marked by dashes indicate the absence of an equivalent counterpart.

Table 3: TBX and OWL in comparison (Gromann, 2015: 108)

| <b>Terminology</b>   | <b>Ontology</b>            |
|----------------------|----------------------------|
| Concept or entry     | Concept or class           |
| ---                  | Individual/Class assertion |
| Generic relation     | Subclass axiom             |
| Partitive relation   | Meronymic object property  |
| Associative relation | (Object property)          |

|                             |                                  |
|-----------------------------|----------------------------------|
| Designation relation        | ---                              |
| Complementary information   | (Annotation property(optional))  |
| Global information          | (Annotation property (optional)) |
| (Datatype textfield)        | Data property                    |
| ---                         | Object property                  |
| ---                         | Axioms                           |
| ---                         | Expressions                      |
| Designation                 | Label (optional)                 |
| Natural language definition | (Annotation property (optional)) |

Three pairs of elements can be regarded as fully corresponding. First, concepts or entries in terminologies correspond to concepts or classes in ontologies. Both structures serve to conceptualize entities within a particular domain. The main difference lies in the language used to represent these conceptualizations: terminologies employ natural language, whereas ontologies rely on formal languages and logical formalisms.

Second, generic relations correspond to subclass axioms. In both cases, a hierarchical relationship is established between a more general concept and a more specific one. However, this correspondence is not universal, as not every hypernym–hyponym relation can be represented accurately as a subclass relation within an ontology.

Third, partitive relations correspond to meronymic object properties. ISO 1087 (2019) defines a partitive relation as a conceptual relation between a comprehensive concept and one of its constituent parts. In ontologies, such part–whole relations are represented by meronymic object properties.

The comparison also reveals an important distinction between the conceptual and linguistic dimensions of the two resources. Ontologies generally treat terms as labels attached to formally defined entities, effectively flatten terms into strings without internal linguistic structure. Terminologies, by contrast, treat terms as structured linguistic units that are consistently used within a specialized domain and whose internal organization is relevant for knowledge representation (Gromann 2015: 109).

Several elements exhibit only partial correspondence. The first case concerns datatype text fields and data properties. In terminological resources, datatypes specify the type of content

permitted within a data category. In OWL, data properties establish relations between individuals and literal values, combining lexical forms or data values with specific datatypes.

A second partial correspondence exists between associative relations and object properties. Both mechanisms represent thematic or semantic associations between two concepts or objects. Nevertheless, object properties possess a considerably broader expressive capacity and are therefore not fully equivalent to associative relations.

Third, complementary information in terminologies corresponds only partially to annotation properties in OWL. Annotation properties are primarily used to provide metadata in OWL 2 about ontology elements and cannot themselves participate in logical relations. Complementary information in terminological resources, however, may be associated with natural-language elements and can be interrelated.

Similarly, global information and annotation properties display only a limited correspondence. Although both provide metadata, no direct equivalence exists between them. In both systems, external resources may be linked to supplement the information contained within the resource itself.

Several syntactic elements have no equivalent counterpart. The first example is the individual/class assertion, which is fundamental in ontology languages but absent from terminological standards. While instances may be represented as concepts in terminology systems, such a representation is semantically inadequate. Ontologies explicitly distinguish between classes and individuals and formally specify class membership through class assertions. No comparable mechanism is defined in ISO standards for terminology management.

A second non-corresponding element is the designation relation. Such relations are particularly important in multilingual terminological resources, where relationships among different designations must be explicitly represented. Ontologies, however, generally do not require formal relations between labels.

Third, object properties have no direct equivalent in terminology standards. In ontologies, object properties establish relationships between individuals and can be constrained through logical characteristics and axioms. Terminologies provide concept relations, but these are limited in number and lack the formal expressiveness of object properties.

The fourth non-corresponding element is represented by axioms. Axioms constitute the logical foundation of ontologies and are expressed in formal languages. Terminological standards do not provide an equivalent mechanism for formal logical reasoning.

Finally, class expressions have no counterpart in terminology systems. In ontologies, class expressions define the conditions that individuals must satisfy to belong to a particular class (Gromann 2015: 110). ISO 1087 (2019) does not specify any analogous construct.

Overall, the syntactic comparison demonstrates that ontologies provide a richer and more expressive formal framework for representing domain knowledge. Terminologies, by contrast, prioritize the organization of concepts and designations for the purpose of specialized communication. While several structural correspondences can be identified, ontologies extend beyond the capabilities of terminological systems through their support for individuals, object properties, logical axioms, and class expressions.

### **2.6.2 Semantic**

The semantic comparison between terminologies and ontologies focuses on the expression of meaning and the representation of real-world entities. Both terminological and ontological resources rely on concepts as the primary means of representing meaning and abstracting knowledge about real-world objects. In this way, both approaches aim to describe and organize knowledge within a specific domain of interest (Gromann 2015: 111).

ISO 704 (2009) declares that concepts are “contextualized in the special language” (ISO 704 2009: 2). Consequently, objects are organized and represented according to the conceptual structures established by a particular specialized community. Ontologies, on the other hand, focus on the formal representation of objects, their classification, and the relationships that exist among them (W3C 2012). While both approaches seek to structure domain knowledge, they differ in their understanding of the nature of concepts and their role in knowledge representation.

Smith et al. (2005) distinguish among several types of concepts discussed in the literature. Concepts employed in ontologies are classified as ontological concepts, which represent abstractions of objects, their attributes, and their properties. Concepts used in terminology are described as epistemic concepts, as they function primarily as units of knowledge within a specialized domain. Smith et al. (2005) further identify psychological

concepts, understood as mental entities, and linguistic concepts, which correspond to the meanings of general-language expressions.

A fundamental difference between terminologies and ontologies concerns the process of concept definition. Terminologies define concepts through their essential and delimiting characteristics expressed in natural language. Ontologies, in contrast, define concepts through necessary and sufficient conditions formulated by means of logical axioms. Consequently, ontological definitions are intended not only to describe concepts but also to support automated reasoning processes.

Another important distinction relates to the assumptions underlying each knowledge representation system. Terminologies generally operate according to a closed-world assumption (Gromann 2015: 112), in which the meaning of a concept is established within a predefined concept system. The focus is therefore placed on lexical semantics and the semantic relations among concepts. Ontologies, on the other hand, adopt an open-world assumption, according to which knowledge can always be extended. Information that is not explicitly represented within an ontology is not considered false; rather, its truth value remains unknown. This principle allows ontologies to accommodate the continuous expansion of knowledge domains without requiring the reformulation of existing structures.

Furthermore, ontologies are grounded in a philosophical understanding of meaning that emphasizes logical consistency and truth conditions. The axioms that structure ontological knowledge must be logically valid and coherent, enabling the inference of new knowledge from existing assertions. Terminologies, by contrast, primarily aim to facilitate human understanding and communication by providing concept definitions and descriptions within specialized domains. As a result, semantic representation in terminologies is oriented toward conceptual clarification, whereas semantic representation in ontologies is designed to support both conceptual modelling and machine reasoning.

### **2.6.3 Pragmatics**

From a pragmatic perspective, the comparison between terminologies and ontologies concerns their respective objectives, intended uses, and development processes. Ontologies are primarily designed to support machine-readable and machine-processable applications, whereas terminologies aim to organize and present knowledge in a manner that is accessible and understandable to human users. Using formal descriptions and logical axioms, ontologies seek

to minimize ambiguity and support automated reasoning. Terminologies, by contrast, reduce misunderstandings in specialized communication by providing standardized representations of domain-specific knowledge.

According to Budin (1996), terminology systems have a regulatory function, making organization a constitutive characteristic of any terminology work. Terminology systems fulfil three principal functions. First, they serve an epistemic function, whereby terminology is interpreted as an instrument for knowledge acquisition and knowledge organisation. Second, they perform an informational function, facilitating the collection, exchange, and management of information, as exemplified by thesauri and terminology database. Third, they fulfil a discourse-oriented function, aiming to optimize of specialized communication by promoting consistency and clarity in the use of domain-specific terms.

Gromann (2015: 113) provides a detailed comparison of the development processes of terminologies and ontologies in order to exemplify their different practical applications. Table 4 below summarize the two workflows recommended by the relevant standardization organizations.

Table 4: Terminology and ontology processes compared (Gromann 2015: 113)

| <b>Terminology</b>                                                                                                                                                                                                                                                    | <b>Ontology</b>                                                                                                                                                                                                                                                                                              |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Identify purpose and scope including a needs analysis of target user group(s)                                                                                                                                                                                         | Identify purpose and scope including target user group(s)                                                                                                                                                                                                                                                    |
| Terminologists analyse discourse (texts) concerning the subject field                                                                                                                                                                                                 | Domain experts identify key vocabulary by means of a shared process                                                                                                                                                                                                                                          |
| Select appropriate terminological mark-up language (TML) based on requirements regarding: <ul style="list-style-type: none"> <li>- planned terminological activity</li> <li>- required data category types and number</li> <li>- intended size of datasets</li> </ul> | Select appropriate profile based on requirements regarding: <ul style="list-style-type: none"> <li>- expressivity</li> <li>- reasoning requirements</li> <li>- size of datasets</li> <li>- importance of scalability</li> <li>- interoperability with, e.g., rule engines or relational databases</li> </ul> |
| Select appropriate syntactic representation and data model description language                                                                                                                                                                                       | Select appropriate syntactic representation language                                                                                                                                                                                                                                                         |
| Specify terminological information: <ul style="list-style-type: none"> <li>- identify concepts, characteristics, and relations in the domain</li> <li>- establish concept diagrams</li> <li>- analyse and modelling concepts</li> </ul>                               | Specify vocabulary using logic: <ul style="list-style-type: none"> <li>- specify namespace</li> <li>- declare objects, concepts, attributes, and relations in the domain</li> <li>- introduce class expressions</li> </ul>                                                                                   |

|                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                  |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"> <li>- introduce hierarchical relations</li> <li>- introduce non-hierarchical relations</li> <li>- define concepts in natural language</li> <li>- specify subject field</li> <li>- specify file description</li> </ul> | <ul style="list-style-type: none"> <li>- introduce class hierarchies</li> <li>- introduce object properties</li> <li>- define concepts by means of axioms (complex classes, etc.)</li> <li>- attribute annotation properties</li> <li>- specify file description and ontology version</li> </ul> |
| Introduce data categories for metadata and data providing additional information                                                                                                                                                                         | Identify objects in domain of interest and make assertions                                                                                                                                                                                                                                       |
| Validation of resulting resource: <ul style="list-style-type: none"> <li>- syntactic check by parsing XML</li> <li>- human validation of content</li> </ul>                                                                                              | Validation of resulting resource: <ul style="list-style-type: none"> <li>- entailment checking</li> <li>- instance retrieval and query checking</li> <li>- consistency checking</li> <li>- human validation of content</li> </ul>                                                                |

The principal differences between the two approaches concern vocabulary specification, population of the resource, validation procedures, and domain modelling. Population refers to the process by which a resource is augmented with terminological or knowledge data, either manually or automatically. In ontology engineering, this process is often supported by automated reasoning mechanisms that facilitate consistency checking and validation. In terminological resources, however, validation generally relies on manual review and expert assessment.

Another significant distinction concerns the specification of domain of interest. Ontologies define concepts through formal axioms expressed in logical languages, while terminologies define concepts through natural-language descriptions and definitions. This difference reflects the distinct approaches to representing reality discussed in the previous sections. Ontologies focus on the formal classification of entities and their membership in classes, thereby supporting reasoning and inference tasks. Terminologies, on the other hand, focus on the management and standardization of terminological information in order to ensure consistency and effectiveness in specialized communication.

In summary, both resources are intended to support knowledge management, but they do so in different ways. Ontologies emphasize the formal representation of knowledge, the modelling of complex relationships, and the automatic derivation of inferences. Terminology, by contrast, concentrate on the description and standardization of concepts and terms, including terminological variation within and across languages, thereby facilitating effective communication in specialized communication.

## 2.6.4 Conclusions and Resulting Implications

Table 5 summarizes the principal similarities and differences between terminologies and ontologies discussed throughout this chapter, drawing primarily on the work of Gromann (2013, 2015). These distinctions are fundamental for establishing the comparative framework adopted in this study, which examines the relationship between manually constructed concept systems and the automatic semantic analysis of texts.

Overall, the comparison demonstrates that terminologies and ontologies should not be regarded as competing approaches but rather as complementary resources. Terminologies contribute linguistic precision, conceptual clarity, and communicative effectiveness, whereas ontologies provide formalization, machine interpretability, and reasoning capabilities. Their integration offers significant potential for bridging human-oriented knowledge organization and machine-oriented knowledge representation. This complementarity is particularly relevant for the present study, which seeks to compare manually developed concept systems with automatically generated semantic structures derived from text analysis. The differences identified in this chapter therefore provide the theoretical basis for evaluating the extent to which ontology learning and semantic technologies can support, enrich, or reproduce traditional terminological representations of specialized knowledge.

Table 5: Summarized comparison of terminologies and ontologies.

| <b>Terminologies</b>                                                        | <b>Ontologies</b>                             |
|-----------------------------------------------------------------------------|-----------------------------------------------|
| Knowledge organization system                                               | Knowledge representation system               |
| Cognitive representation of reality and its graphical representation        | Real-world objects                            |
| Cognitive, communicative and informational representation of the real world | The being itself                              |
| Domain-specific concepts designed by terms in a specialized language        | Formal and explicit knowledge representations |
| Closed-world resource                                                       | Open-world resource                           |
| Treat concepts and terms                                                    | Refer and relate to objects                   |
| Natural language                                                            | Formal languages                              |
| TBX                                                                         | OWL                                           |
| Use terms consistently in domain                                            | Terms flattened into strings                  |
| Conceptualization of a specific domain                                      | Conceptualization of ontology classes         |
| Epistemic concepts                                                          | Ontological concepts                          |

|                                                        |                                                          |
|--------------------------------------------------------|----------------------------------------------------------|
| Concepts designed by terms in natural language         | Concepts defined by axioms                               |
| Concepts grouped by properties                         | Concepts grouped by classes                              |
| Essential and delimiting characteristics               | Necessary and sufficient characteristics                 |
| Lexical semantics and semantic relations               | Philosophical notion of meaning                          |
| Abstracts away from objects                            | Abstract, describe and group objects by ontology classes |
| Specialized communication optimized                    | Reasoning task                                           |
| Defining concepts and their interrelations             | Interpretation by machine.                               |
| Terminological differences within and across languages | Complex relations between objects                        |
| Entry                                                  | Class                                                    |
| Generic relation                                       | Subclass axiom                                           |
| Partitive relation                                     | Meronymic object property                                |
| Associative relation                                   | Object property                                          |

### 3. Related Work

In this master's thesis a process has been explored, which mirrors the workflow of Gromann (2020), by investigating the differences and similarities between ontologies and terminologies, to determine, if ontologies can be applicable as effective support for terminologies. The central core of both this master thesis and Gromann (2020) is the use of concept systems: in terminologies are used to structure and organize concepts, while they enrich ontologies with linguistic specification.

Gromann (2020) discusses multiple benefits associated with the use of concept systems. Primarily, concept systems represent a solid basis for consistent definitions, as well as effective approach for reducing the ambiguity in definitions themselves. In fact, the integration of ontologies with terminological systems facilitates the elimination and correction of duplicates. In this respect, concept systems can be proven useful as quality check tool in terminology work. Furthermore, they foster easier comparison between different multilingual terminologies. Ultimately, concept systems enable a clearer understanding and production of specialized texts, by enriching domain-specific knowledge.

However, concept systems are not used systematically due to two main limitations: the lack of automated tools and the resultant high cost and time involved for manual approaches. Consequently, specialized texts constitute one of the most valuable resources for both the manual and automatic creation of concept systems. Formal knowledge representation languages enable the flexible specification of semantic relations, while ontology learning techniques facilitate the automatic extraction of ontological structures from natural language texts. These developments offer promising opportunities for reducing the effort involved in concept-system construction and for strengthening the integration of terminological and ontological approaches to knowledge organization.

Eventually Gromann (2020) underlines also that the logic and linguistic representation of knowledge of terminologies and ontologies are intrinsically different, and the two resources combined lose both part of their functionalities. This paradox is defined as ontology-epistemology divide (Bodenreider et al. 2004): terminologies lose epistemic knowledge, while ontologies lose the ability of logical relations and the logical reasoning, when combined in an ontology-terminology model. In the final analysis, a terminological concept system cannot be substituted by an ontology.

Three different onto-terminological models can be identified in Gromann (2020). The first model represents the terminological information as an ontology (Lacasta et al. 2010), defined in bibliography as terminological ontologies. However, this solution emphasises the ontology-epistemology divide. The second model integrates an ontological structure with a terminological approach. This model can be conducted through two different solutions. First, terms are modelled as instances of ontological classes (Reymonet et al. 2011). Second, terminological principles are explored for ontology design (Aussenac-Gilles et al. 2008). In both cases, terminological information is used as metadata.

The third model, known as Ontoterminology (Roche et al. 2009), links terminology to a formal ontology, where the terminology is preserved while it is integrated into an ontology. A related method and its application, T-Mint, is developed in Gromann (2014). This approach relies on the principle of semantic by reference (Cimiano et al. 2009: 431.), whereby the terminological entry is linked to an ontological class and refers to an external linguistic resource. In fact, the meaning of a terminological entry is specified in relation to the ontology, as a result, the terminology is endowed with formal semantics. This permits the representation of multiple relations, including ontological relations, terminological concept relations, relations between terminological entries, and reference relationships between ontology and terminology. Reference relations enable the logical representation of a concept's meaning under an ontological point of view, while lexical-semantic, pragmatic and syntactic information is represented under a terminological point of view.

Using ontologies as terminological concept systems offers concrete advantages. First, unstructured sources can be automatically processed to transform them into machine-readable knowledge, facilitating data extraction and organization. Eventually, the ontology learning from texts is employed to extract terminological concept systems, thereby structuring the information. The separate concept representation in onto-terminological reference can resolve the problem of terminological gaps and linguistic nuances, improving clearness and precision. Moreover, automatic reasoners verify the concept consistency, by validating the linguistic content in accordance with terminology criteria. The use of query languages, such as SPARQL, facilitate the search of concepts and enable an efficient navigation of ontology. Lastly, ontologies can be enriched with linguistic specifications.

At the same time, the lack of a correspondence between ontology and terminology represents a disadvantage. In fact, in case the terminology defines a concept that is not included

in the ontology, it is necessary to integrate manually the missing element in the ontology. For this reason, the ontology learning and terminology work from domain-specific texts represent an interesting approach (Gromann 2020).

To conclude, ontologies can be used as a semantic support for the terminological concept systems. Term extraction combined with ontology learning methods is an interesting application of this idea (Gromann 2020). In fact, the integration of both resources could improve the usability of ontology used as semantic support for terminologies. Both resources will benefit from this cooperation: terminologies will be integrated with formal semantics, while ontologies will acquire linguistic richness.

In the last years various applications of ontologies have been developed in the medical field, with also the scope of semantic interoperability. In Mora et al. (2024) this is deeply analysed through the medicine paradigm Personalized, Preventive, Predictive, Participative and Precision (P5). In this paperwork particular attention is given to the evolution of terminology over time, thereby confirming the Socio-Cognitive Theory of Terminology, which assigns a central role to the gradual change in the meaning of categories, concepts, and their designations as human knowledge develops. To this end, the use of the CTS2 (OMG) is validated as a means of transforming OWL ontologies while preserving their structure and supporting the diachronic evolution of the terminological system.

Fung and Bodenreider (2023) analyse the main medical ontologies, which have proven to be most relevant in clinical research, their use and their benefits. The distinction between research metadata ontologies and data content ontologies is significant. The former's main purpose is simplifying the automated management of clinical trials, while the latter's purpose is the standardization of data and semantic interoperability. Data content ontologies, in fact, design the information model and data elements, while research metadata ontologies model the data obtained during academic analysis; therefore, they illustrate the standard for formal representation of research.

The ontology layer was informed by several well-established biomedical research ontologies, including the Ontology of Clinical Research (OCRe), the Ontology for Biomedical Investigations (OBI), and the Biomedical Research Integrated Domain Group (BRIDG) model. OCRe provides a formal representation of clinical study design and protocol semantics, supporting the full lifecycle of clinical research (Sim et al. 2014). OBI offers a comprehensive ontology for describing biomedical investigations, encompassing study planning, execution,

and reporting across diverse experimental domains (Bandrowski et al. 2016). BRIDG serves as a domain analysis model that captures the semantics of protocol-driven biomedical research and regulatory processes, facilitating interoperability among clinical research information systems (Fridsma et al. 2008). Unlike OCRE and OBI, BRIDG is often described as a domain analysis model rather than a formal OWL ontology, although it is frequently used alongside biomedical ontologies and ontology-based frameworks.

Another important issue in the literature is the limited interoperability among biomedical terminologies and ontologies, resulting from their development by different institutions and research communities. To address this challenge, several integrative resources and repositories have been established, including the Unified Medical Language System (UMLS) and BioPortal. These platforms provide mechanisms for mapping, integrating, and accessing heterogeneous biomedical vocabularies and ontologies, thereby promoting semantic consistency and interoperability (Bodenreider 2004; Noy et al. 2009). The widespread adoption and consistent use of such resources can support the expansion of standardized, integrated, and reusable ontology-based approaches in biomedical research.

## 4. Research Methodology

The proposed research methodology can be described as qualitative empirical research, based on a case study and inductive approach. The research process is structured into three principal stages: the preparatory phase, the Python-based automated comparison of two sources (terminologies and ontologies) and the manual analysis of identified matches. Each of these steps will be examined in detail in the following paragraphs.

The preparatory phase plays a crucial role in research process. This phase is constituted by the identification of text sources, the manual creation of terminological resources using SDL MultiTerm, and the automated construction of ontologies via FRED.

Firstly, ten sample texts (see Paragraph 9.1, Text Sources) were selected from English-language Wikipedia. The analysed texts cover different topics within the medical domain. The primary aim of the research is the comparison of conceptual models, namely an ontological model and a terminological model, and the identification of correspondences between them. More specifically, this empirical study investigates whether the concepts, their designations, and their relations, as derived from the input texts, are comparable across the two conceptual systems. The selected domain is medicine, and the ten texts represent distinct areas within this field:

- Text 1: Diabetes Mellitus
- Text 2: Multiple Sclerosis
- Text 3: Colonoscopy
- Text 4: Caesarean Section
- Text 5: Nonsteroidal anti-inflammatory drugs
- Text 6: Hormonal contraception
- Text 7: Antidepressant
- Text 8: Gastroparesis
- Text 9: Inflammation
- Text 10: Infection

Secondly, a manually curated concept system in TBX format was created for each text. This was achieved by using the Term Base function in MemoQ, exporting the data in XML format, and subsequently converting it into TBX using SDL MultiTerm. The terminology was developed following a translator-oriented approach, whereby terminological work is

undertaken as a preparatory stage of the translation process. In this sense, each text was read and analysed with the aim of understanding the domain of interest and extracting relevant terms, which were subsequently associated with their definitions. By contrast, FRED identifies and analyses only those concepts explicitly present in the source texts, establishing the relationships among them. Consequently, lexical variants such as synonyms (when not systematically attested in the texts) are not detected by the system. Such elements can therefore be identified only through manual analysis or when they explicitly occur in the textual data.

Term-level information lists terms and their synonyms. Language-level information lists only the two categories definition and source of the definition. Considering that FRED exploits the Wikipedia hyperlink structure when analysing the texts, it has been decided to quote Wikipedia web pages for definitions of each concept. In one instance, Wikipedia was not consulted as a source, since no standalone entry for *gastric emptying* could be identified at the time of consultation (July 2021).

The third stage consists of an online, automated semantic analysis of the texts using FRED, an ontology learning tool. Each text is analysed individually by FRED. The fourth and final stage involves a comparison between the FRED output and the manually curated concept systems. This comparison is conducted first through an analysis of the resulting statistical data and subsequently through a qualitative examination of the conceptual relations identified in both sources.

#### **4.1 Production of Manually Created Concept System**

First, the ten sample texts were imported as individual documents into MemoQ, and automatic term extraction was performed using the Extract Terms function. At this stage, the tool generates a list of candidate terms based on frequency and linguistic patterns. These candidates are subsequently reviewed by the translator and either validated or discarded. During this process, the stop word list can be extended in order to exclude non-relevant lexical items such as articles, prepositions, or adverbs. The resulting candidate terms were exported as an Excel file for further processing.

For each text, a separate term base was created in order to ensure clear separation between the different source texts and their respective terminological inventories. The relevant data categories for this study were defined as term, definition, and source. These categories are sufficient for the purpose of this research, and additional fields were not considered.

Term-level information lists terms and their synonyms, while language-level information lists only the two categories definition and source of definition. Considering that FRED relies on the Wikipedia hyperlink structure for semantic analysis of texts, it has been decided to quote Wikipedia web pages for terms definitions.

Thirdly, the term bases were exported as XML files. Since TBX files can be imported into MemoQ but not exported directly, each term base was exported in XML format, along with the associated XDT and XDL files. The XML files were then imported into SDL MultiTerm in order to convert them into TBX format. TBX is the standard format for terminology exchange (ISO 30042:2019) and is therefore used as the reference format in terminology management workflows.

## 4.2 Editing of TBX File in SDL Multiterm

As the term bases are imported into SDL MultiTerm, the subsequent step involves their editing. This fundamental phase is necessary to enable the annotation of concept relations. Therefore, it is also crucial for the subsequent comparison between TBX files and Turtle files. Each term base is then edited to incorporate the concept relations into the entry structure at the entry level. The concept relations employed are summarized in Table 6.

Table 6: Concept relations employed in the study

| Concept relation  | Description                                                                                                                                                                            | Linguistic realization                      |
|-------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------|
| genericRelation   | The relation between a specific concept (subordinate) to a generic one (superordinate), it represents an abstraction relation, a typical definition relationship.                      | is a                                        |
| partitiveRelation | The relation between the part to the comprehensive concept.                                                                                                                            | is part of                                  |
| activityRelation  | The relation from actor to activity, from actor to object or from activity to result.                                                                                                  | is made by; is carried out by; is made with |
| spatialRelation   | The concept is in a spatial relation because of the relative physical position in relation to another concept. It also describes the relation from the moving object to the fixed one. | it is located                               |
| temporalRelation  | A concept stands in a temporal relationship to                                                                                                                                         | before; later                               |

|                       |                                                                                                                                                                                                  |                                     |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------|
|                       | another term by virtue of chronological events.                                                                                                                                                  |                                     |
| causalRelation        | The relation between cause and effect, in order to describe the causal connection between the concepts. One concept represents the cause, the other the consequence.                             | causes                              |
| developmentalRelation | It describes the evolution through different stages; in fact, the concept is related to the other one, as chronological consequence. the previous state to the later                             | becomes                             |
| instrumentalRelation  | The relation between a goal (or activity) and the instrument used to achieve it.                                                                                                                 | by means of; using; in order to     |
| originationRelation   | A concept has an origination relation to another concept. It describes the relation between product and its origins, a product and its producer, a product and the associated production method. | comes from                          |
| ownershipRelation     | A concept (an object) is the property of another one (the owner).                                                                                                                                | belongs to                          |
| propertyRelation      | The relation between an object and its properties or attributes; it often describes materials and their characteristics.                                                                         | has properties, is characterized by |
| associativeRelation   | This concept relation has to be used only when the relation between two concepts cannot be described by another relations.                                                                       |                                     |
| (synonymyRelation)    | This relation denotes the relationship between designations that refer to the same concept (ISO 1087 2019)                                                                                       |                                     |

The final relation reported in Table 6, namely the synonymy relation, is shown in parentheses because it was not included in the analysis.

After importing the XML files, it is required to edit the definition of the term bases. As the term bases are imported from MemoQ and all entries are processed, it is possible to modify

the structure. Each concept relation can be added manually one by one in the definition of structure.

The concept relations are manually added to the entry level of the project in SDL MultiTerm, resulting in the data categories depicted in Figure 19.

| Entry Structure       |           |          |
|-----------------------|-----------|----------|
|                       | Mandatory | Multiple |
| <b>Entry level</b>    |           |          |
| temporalRelation      |           | •        |
| synonymyRelation      |           | •        |
| spatialRelation       |           | •        |
| propertyRelation      |           | •        |
| partitiveRelation     |           | •        |
| ownershipRelation     |           | •        |
| originationRelation   |           | •        |
| instrumentalRelation  |           | •        |
| genericRelation       |           | •        |
| developmentalRelation |           | •        |
| causalRelation        |           | •        |
| associativeRelation   |           | •        |
| activityRelation      |           | •        |
| Client                |           |          |
| Domain                |           |          |
| ID                    |           |          |
| Note                  |           |          |
| Project               |           |          |
| Subject               |           |          |
| Image                 |           |          |
| <b>Index level</b>    |           |          |
| Definition            |           |          |
| Forbidden term        |           | •        |
| <b>Term level</b>     |           |          |
| Usage example         |           |          |
| Part of speech        |           |          |
| Number (grammar)      |           |          |
| Gender (grammar)      |           |          |
| Definition            |           |          |

Figure 19: SDL Multiterm: Entry Structure Definition

Before exporting the term bases as TBX files, it is necessary to encode the concept relations within each entry structure. Each concept may be associated with one or more concept relations. To annotate these relations, the concept is selected, the appropriate concept relation is assigned, and the identifier (ID) of the related concept is specified as its value.

An example is provided in Figure 20, where the relation between the concepts identified by ID 26 and ID 18 is a generic relation. This is represented by assigning the category *genericRelation* to the concept with ID 26 and indicating 18 as the relation value, thus establishing a link from a specific concept to a generic one. Once this phase is completed, the term bases are exported as TBX files.

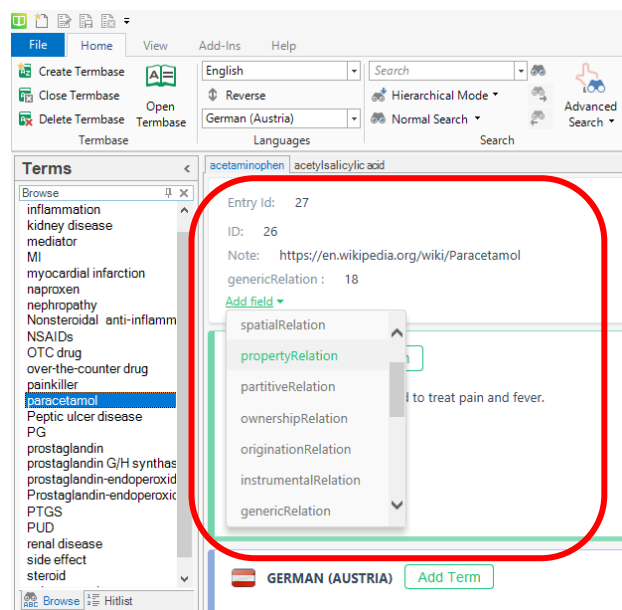


Figure 20: SDL Multiterm: establishing of concept relations between entries.

### 4.3 Online Semantic Analysis with FRED

Each text is analysed using FRED, a tool for automatic semantic annotation that extracts RDF/OWL representations from natural language. FRED is employed in this study as it enables the consistent and formalised transformation of textual data into structured knowledge, thereby facilitating comparison with TBX representations and supporting interoperability across formats.

The process of analysing the selected texts with FRED via its API is illustrated through screenshots. With the exception of the *Response content format* parameter (set to *text/turtle*), the default settings were retained throughout the analysis. The natural language text is entered into the input field and processed automatically, and the resulting output is subsequently stored for further analysis. In Figure 21 is reported the screenshot of FRED settings for online analysis.

During the analysis of the sample texts, minor adjustments were required to ensure meaningful results. Specifically, in Texts 6 and 8, the symbol “%” was replaced with the lexical form “percent”, and parenthetical content was removed.

default

GET /stlab-tools/fred

Generate RDF from natural language text.

Parameters

Name

Description

Authorization
required

The authorization bearer. Type "Bearer xxx-yyy-zzz", where is your secret token.

string

(header)

Bearer 48f464fb-211f-3801-a0bb-3ed68f026de7

text
required

The input natural language text.

string

(query)

text - The input natural language text.

prefix

The prefix used for the namespace of terms introduced by FRED in the output. If not specified fred: is used as default.

string

(query)

prefix - The prefix used for the namespace of terms

namespace

The namespace used for the terms introduced by FRED in the output. If not specified http://www.ontologydesignpatterns.org/ont/fred/domain.owl# is used as default.

string

(query)

namespace - The namespace used for the terms in

wsd

Perform Word Sense Disambiguation on input terms. By default it is set to false.

boolean

(query)

--

wfd

Perform Word Frame Disambiguation on input terms in order to provide alignments to WordNet synsets, WordNet Super-senses and Dolce classes. By default it is set to false.

boolean

(query)

--

wfd\_profile

The profile associated with the Word Frame Disambiguation

string

(query)

b

tense

Include temporal relations between events according to their grammatical tense. By default it is set to false.

boolean

(query)

--

roles

Use FrameNet roles into the resulting ontology. By default it is set to false.

boolean

(query)

--

textannotation

The vocabulary used for annotating the text in RDF. Two possible alternatives are available, i.e. EARMARK and NIF.

string

(query)

earmark

semantic-subgraph

Generate a RDF which only expresses the semantics of a sentence without additional RDF triples, such as those containing text spans, part-of-speeches, etc. By default it is set to false.

boolean

(query)

--

Execute

Responses

Code

Description

200

The RDF produced by FRED for the text provided as input.

Response content type

application/rdf+xml

application/rdf+xml

text/turtle

application/rdf+json

text/rdf+n3

text/rdf+nt

image/png

Figure 21: FRED: settings

#### 4.4 Comparison Method

The comparison method was developed after the creation of the term bases and the generation of the automatic semantic analysis produced by FRED. In order to compare the two different types of resources, namely the manually curated SDL terminology resource and the semantic output generated by FRED, both datasets were parsed and analysed according to a set of predefined criteria. The comparison model was designed to evaluate the extent to which FRED is capable of automatically extracting concepts and semantic relations from specialised texts and to assess the degree of correspondence between the automatically extracted information and the manually curated terminology resource. The methodology therefore combined

quantitative and qualitative analyses in order to evaluate both the amount of extracted information and its semantic reliability.

The quantitative analysis was divided into two principal dimensions. The analysis of terms and the analysis of semantic relations. In the first phase, the comparison focused on the extraction of concepts and terminology units. Several parameters were selected in order to evaluate the performance of FRED in identifying domain-specific terminology.

The first parameter considered was the number of terms identified in FRED. This value corresponds to the total number of lexical or conceptual units automatically extracted by the semantic parser from the analysed corpus. The parameter provides an initial indication of the productivity and extraction capacity of the automatic system. However, the absolute number of extracted terms alone does not provide sufficient information regarding the quality or relevance of the identified concepts. The second parameter was the number of terms contained in SDL. This dataset represents the manually curated terminology resource and therefore constitutes the reference model against which the automatic extraction process was evaluated. Since SDL was manually compiled according to terminological principles, it was assumed to provide a reliable representation of the relevant domain terminology.

A further parameter consisted of the number of common terms shared by FRED and SDL. This measure identifies the overlap between the automatically extracted concepts and the concepts already present in the manually curated terminology base. The greater the overlap between the two resources, the greater the capability of FRED to identify terminology that is considered relevant within the domain.

In addition to the numerical overlap, a list of common terms was generated. This list enabled a more detailed qualitative inspection of the matching terminology units and facilitated the identification of recurring patterns, domain-specific terminology, synonymous forms, or possible extraction inconsistencies. The examination of common terms was particularly important in order to understand whether the extracted concepts corresponded semantically to the manually curated terminology and whether terminological variation affected the matching process.

To quantify the similarity between the two sets of terms, the Jaccard index (Jaccard 1901) was employed. The Jaccard index is a statistical coefficient used to measure similarity between two finite sample sets. In this study, the index was calculated by comparing the set of terms extracted by FRED with the set of terms contained in SDL. More specifically, the

coefficient was obtained by dividing the number of shared terms by the total number of unique terms contained in both datasets. The resulting value ranges from 0 to 1, where 0 indicates no overlap and 1 indicates complete correspondence between the two resources. The purpose of applying the Jaccard index was to provide a measure of terminological similarity that would allow a more objective comparison independently of the size of the datasets.

The evaluation of extracted terms was further conducted using standard information retrieval metrics, namely precision, recall, and F-score. These metrics are commonly employed in computational linguistics and information extraction tasks in order to assess system performance.

Precision measures the proportion of relevant terms among all terms extracted by FRED (Powers 2007). In the context of this study, precision indicates how many of the automatically extracted terms correspond to valid terminology units present in the SDL resource. A high precision value therefore indicates that the automatic extraction process produces relatively few incorrect or irrelevant terms.

Recall measures the proportion of relevant terms successfully retrieved by FRED with respect to the total number of relevant terms present in SDL (Powers 2007). This metric evaluates the completeness of the extraction process by indicating how effectively the system identifies the terminology contained in the reference resource. High recall values indicate that the system succeeds in identifying a large proportion of the relevant terminology.

The F1-score, or harmonic mean of precision and recall, was calculated in order to provide a balanced measure of extraction performance. Since precision and recall often exhibit an inverse relationship, the F-score makes it possible to evaluate the overall effectiveness of the extraction process by considering both correctness and completeness simultaneously.

A comparable methodology was applied to the analysis of semantic relations. Semantic relations constitute a fundamental component of terminology resources because they represent conceptual connections between terms and contribute to the semantic structure of the domain knowledge.

The first parameter analysed was the number of relations identified by FRED. This value corresponds to the total number of semantic relations automatically generated by the system, including hierarchical, associative, or predicate-based relations extracted from the corpus.

The second parameter was the number of relations contained in SDL. As in the case of terminology units, the manually curated SDL relations served as the reference dataset against which the automatic semantic extraction process was evaluated.

The number of common relations represented the overlap between the semantic relations identified automatically by FRED and those encoded in SDL. This parameter made it possible to evaluate whether the automatic system was capable not only of identifying concepts but also of reconstructing meaningful semantic connections between them.

A list of common relations was additionally generated in order to support the qualitative inspection of matching semantic structures. The analysis of common relations enabled the identification of correspondences and divergences between the two approaches and facilitated the examination of relation types recognized by FRED.

The similarity between the two sets of semantic relations was also measured through the Jaccard index. In this context, the coefficient quantified the proportion of shared semantic relations relative to the total number of distinct relations identified across both datasets. This measure provided an overall indication of semantic correspondence between the manually curated resource and the automatic semantic analysis.

Another parameter considered in the evaluation was accuracy. According to ISO 5725-1 (1994), accuracy measures the proximity of an analysed value to the true value. In information extraction systems, accuracy is generally defined as the ratio between correctly classified instances, including true positives and true negatives, and the total number of analysed instances. Within this study, accuracy was used to evaluate the reliability of the automatic semantic extraction process in correctly identifying valid semantic relations.

Finally, semantic relations were evaluated using precision, recall, and F-score metrics. Precision measured the proportion of correctly identified semantic relations among all relations extracted by FRED, whereas recall measured the proportion of relevant relations successfully retrieved from the SDL reference resource. The F-score provided a combined measure of semantic extraction performance by balancing correctness and completeness.

Following the quantitative analysis, a qualitative analysis was conducted in order to examine the semantic nature and structural characteristics of the extracted relations. This phase was necessary to determine whether the relations identified by FRED corresponded conceptually and functionally to those encoded in the manually curated terminology resource.

Particular attention was paid to the consistency of semantic relation types, the adequacy of conceptual associations, and the presence of semantic ambiguities or extraction errors.

The integration of quantitative and qualitative analyses ultimately enabled the identification of the approach that produced a greater quantity of information as well as the approach that provided higher-quality semantic representations. Consequently, the methodology made it possible to evaluate the suitability of each approach according to the specific research objectives, textual characteristics, and domain-specific terminology involved in the study.

## 5. Results

In the current chapter, the findings of the study are divided into quantitative and qualitative results. First, the results are presented objectively, followed by an analysis of the matching relations and the most interesting cases retrieved by FRED.

### 5.1 Quantitative Findings

The following results are presented objectively by evaluating the correspondence between the terms and relations extracted by FRED and those identified through SDL analysis. The comparison was conducted by analysing whether the concepts and relations extracted by FRED corresponded to the same or semantically similar terms and relations examined during the SDL analysis, i.e. the terminology work. In the case of matching relations, the evaluation considered not only the presence of a relation but also the kind of relation. The degree of overlap between the two datasets was measured through quantitative evaluation metrics calculated in Python. These metrics included the Jaccard index, precision, recall, accuracy, and F1-score.

The Jaccard index was employed to measure the degree of overlap between the items extracted by FRED and those identified through SDL analysis. Precision measures the proportion of concepts or relations extracted by FRED that also appear in SDL, while recall measures the proportion of SDL concepts or relations successfully retrieved by FRED. The F1-score is defined as the harmonic mean of precision and recall and provides a balanced evaluation of extraction performance. Accuracy was also considered because the F1-scores were consistently low, mainly due to the large discrepancy between the number of elements extracted by FRED and those identified in SDL. Consequently, the inclusion of Accuracy provided additional information regarding the overall correspondence between the two datasets.

Across all ten analysed texts, the average F1-score for term extraction was 28.92%, whereas the average F1-score for relation extraction was considerably lower, with a value of 2.15%.

Table 7: Comparison between extracted terms in FRED and terms in SDL.

| <b>Text ID</b> | <b>Terms in FRED</b> | <b>Terms in SDL</b> | <b>Number of common terms</b> | <b>Jaccard coefficient (%)</b> | <b>Precision (%)</b> | <b>Recall (%)</b> | <b>F1-score (%)</b> |
|----------------|----------------------|---------------------|-------------------------------|--------------------------------|----------------------|-------------------|---------------------|
| TEXT 1         | 190                  | 61                  | 21                            | 18.10                          | 21.88                | 51.22             | 30.66               |
| TEXT 2         | 124                  | 57                  | 11                            | 14.20                          | 19.20                | 35.29             | 24.87               |
| TEXT 3         | 156                  | 57                  | 13                            | 15.23                          | 19.11                | 42.86             | 26.43               |
| TEXT 4         | 142                  | 43                  | 12                            | 19.39                          | 22.54                | 58.18             | 32.49               |
| TEXT 5         | 126                  | 55                  | 12                            | 13.53                          | 18.25                | 34.33             | 23.83               |
| TEXT 6         | 167                  | 29                  | 8                             | 10.27                          | 11.38                | 51.35             | 18.63               |
| TEXT 7         | 130                  | 57                  | 14                            | 16.86                          | 22.31                | 40.85             | 26.86               |
| TEXT 8         | 182                  | 78                  | 22                            | 22.61                          | 28.57                | 52.00             | 36.88               |
| TEXT 9         | 169                  | 60                  | 11                            | 18.23                          | 21.89                | 52.11             | 30.83               |
| TEXT 10        | 139                  | 70                  | 26                            | 21.76                          | 30.22                | 43.75             | 35.74               |

Table 7 presents the comparison between the terms extracted by FRED and the terms identified through SDL analysis. For each source text, the table reports the total number of terms extracted by FRED, the number of terms identified in SDL, the number of common terms between the two datasets, and the corresponding evaluation metrics.

The data show that FRED consistently extracted a higher number of terms than those identified through SDL analysis. For example, in Text 1, FRED extracted 190 terms, whereas SDL identified 61 terms, with 21 common terms between the two datasets. This resulted in a Jaccard index of 18.1%, a precision of 21.88%, a recall of 51.22%, and an F1-score of 30.66%. Similar tendencies can be observed across the remaining texts.

The highest correspondence between the two datasets was observed in Text 8, which achieved the highest Jaccard index (22.61%), the highest recall (52.00%), and the highest F1-score (36.88%). In this case, 22 common terms were identified between 182 terms extracted by FRED and 78 terms identified in SDL. Similarly, Text 10 achieved the highest precision value (30.22%), indicating that nearly one-third of the concepts extracted by FRED corresponded to concepts identified manually.

The lowest performance was observed in Text 6, which obtained the lowest Jaccard index (10.27%), the lowest precision (11.38%), and the lowest F1-score 18.63%. Although recall remained relatively high (51.35%), only eight common terms were identified between the two datasets. Overall, recall values were consistently higher than precision values across all texts.

Table 8: Comparison between analysed relations by FRED and manually reported relations in SDL

| Text ID | Relations in FRED | Relations in SDL | Number of common relations | Jaccard index (%) | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---------|-------------------|------------------|----------------------------|-------------------|--------------|---------------|------------|--------------|
| TEXT 1  | 529               | 29               | 10 (5)                     | 2.12              | 34.48        | 2.24          | 28.57      | 4.16         |
| TEXT 2  | 261               | 31               | 2 (1)                      | 0.80              | 6.45         | 0.90          | 6.25       | 1.58         |
| TEXT 3  | 301               | 36               | 2 (1)                      | 0.68              | 5.56         | 0.77          | 5.41       | 1.35         |
| TEXT 4  | 326               | 26               | 9 (4)                      | 2.86              | 34.62        | 3.07          | 29.03      | 5.56         |
| TEXT 5  | 266               | 30               | 2 (1)                      | 0.78              | 6.67         | 0.87          | 6.45       | 1.54         |
| TEXT 6  | 482               | 21               | 0                          | 0.00              | 0.00         | 0.00          | 0.00       | 0.00         |
| TEXT 7  | 284               | 37               | 2 (1)                      | 0.72              | 5.41         | 0.83          | 5.26       | 1.43         |
| TEXT 8  | 384               | 44               | 6 (3)                      | 1.58              | 13.64        | 1.78          | 12.77      | 3.12         |
| TEXT 9  | 412               | 35               | 2 (1)                      | 0.51              | 5.71         | 0.56          | 5.56       | 1.01         |
| TEXT 10 | 351               | 44               | 3 (2)                      | 0.87              | 6.82         | 0.99          | 6.52       | 1.72         |

Table 8 presents the comparison between the relations extracted by FRED and the relations identified through SDL analysis. The table reports the total number of relations extracted by FRED, the number of relations identified in SDL, the number of common relations, and the corresponding evaluation metrics. In the column “Number of common relations”, the number of unique relations is reported in brackets, since several matching relations correspond to duplicated instances generated by FRED. This pattern is analysed further in Section 5.2 and proves important for a closer qualitative inspection.

The figures in Table 8 suggest a substantially lower correspondence for relation extraction than for term extraction. FRED generated a very large number of relations, ranging from 261 relations in Text 2 to 529 relations in Text 1, whereas the number of SDL relations varied between 21 and 44. In Text 1, FRED extracted 529 relations, whereas SDL identified 29 relations, with 10 common relations between the two datasets. This produced a Jaccard index of 2.12%, an accuracy of 34.48%, a precision of 2.24%, a recall of 28.57%, and an F1-score of 4.16%. Thus, Text 1 also demonstrated comparatively stronger performance.

The values remain consistently low across the analysed texts. Several texts, including Texts 2, 3, 5, 7, 9, and 10, reported F1-scores ranging between 1.01% and 1.72%, indicating a minimal overlap between the extracted and analysed relations. While Text 1 reports a F1-score of 4.16% while Text 8 3.12%.

The highest correspondence was observed in Text 4, which achieved the highest F1-score (5.56%) and the highest accuracy (34.62%). Nine common relations were identified between the 326 relations extracted by FRED and the 26 relations identified through SDL analysis.

At the opposite end of the spectrum, Text 6 showed no overlap between the two datasets. No common relations were identified, resulting in Jaccard Index, Precision, Recall, Accuracy, and F1-score values of 0%. This represents the lowest performance among all analysed texts.

Precision values remained below 4% in all cases, demonstrating that only a very small proportion of the relations generated by FRED corresponded to those identified in SDL. Relation extraction showed substantially lower correspondence than term extraction.

Overall, term extraction achieved substantially higher correspondence than relation extraction. Average recall exceeded 46% for terms, whereas relation precision remained below 4% for all texts. These findings indicate that FRED is considerably more effective at identifying

domain concepts than reproducing the conceptual structures established through manual terminology analysis.

## 5.2 Qualitative Findings

Following the quantitative comparison, a qualitative examination of the shared relations was conducted. The analysis revealed that several of the common relations identified between FRED and SDL corresponded to duplicate relation instances. The present section examines the common relations identified for each text and explores the conceptual similarities and differences between the two approaches.

As reported in Table 8, several matching relations identified during the quantitative comparison proved to be duplicates upon closer inspection. In Text 1, ten common relations were identified, five of which are duplicates. In Text 2, Text 3, Text 5, Text 7 and Text 9, two common relations were found, both representing duplicate instances. In Text 4, four of the nine shared relations are duplicates. Text 8 presents six shared relations, of which three are duplicates, while Text 10 presents three shared relations, including one duplicate. By contrast, no common relations were identified in Text 6. Therefore, the qualitative analysis focuses primarily on the differences between the relations identified by the two approaches.

For the purposes of this study, a comprehensive relation-by-relation comparison of the FRED output for each text is beyond the scope of the analysis. Nevertheless, the matching relations for each text are first reported, after which representative examples are discussed to illustrate the similarities and differences between the two approaches. Importantly, these differences do not necessarily reflect incorrect interpretations by either approach. Rather, they reveal distinct conceptualisation strategies and modelling approaches.

In Text 1 ten common relations are reported, but only five unique matches are found. This leads to a 50% duplication rate and also explains why accuracy is relatively high (34.48%) while precision remains extremely low (2.24%). The unique matching relations are diabetes → cause → complications, cells → respond → insulin, insulin-dependent diabetes mellitus → subClassOf → diabetes mellitus, juvenile diabetes → subClassOf → diabetes and adult-onset diabetes → subClassOf → diabetes. Nevertheless, Text 1 reveals two noteworthy cases. First, the concept metabolic disorder (ID 1 in SDL) appears in both sources; however, only the manually encoded terminology in SDL reports the associative relation with metabolism (ID 2 in SDL). In the FRED output, this relation is absent because the term metabolism does not explicitly occur in the source text. On the other hand, FRED identifies the synonymic relation

between hyperglycemia (ID 3) and high blood sugar level through the relation owl:equivalentClass:

```
<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#HighBloodSugarLevel>
    a          <http://www.w3.org/2002/07/owl#Class> ;
    <http://www.w3.org/2000/01/rdf-schema#subClassOf>

<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#BloodSugarLevel> ;

<http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#associatedWith>

<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#HighBloodSugar> ;
    <http://www.w3.org/2002/07/owl#equivalentClass>

<http://dbpedia.org/resource/Hyperglycemia> .
```

Text 2 reports only one unique relation, as both reported matches are identical. FRED classifies viral infection as a subtype of infection through the relation type, thereby establishing a taxonomic hierarchy. In SDL, by contrast, the connection is represented through a causalRelation, reflecting the fact that the infection is caused by a virus.

```
{'Fred.': [['infection'], 'type', ['viral infection', 'Viral Infection']], 'SDL: ': [['infection'],
'causalRelation ', ['viral infection', 'viral disease', 'infectious disease']]}
```

In this case, FRED interprets the infection as a type of viral infection, while in SDL the infection results from a viral infection. While both relations are semantically plausible, they encode different dimensions of the conceptual structure.

Text 3 presents only the unique relation, i.e. colon cancer  $\rightarrow$  subClassOf  $\rightarrow$  cancer. This relationship represents one of the strongest correspondences observed in the whole dataset. In the manual terminology of SDL, the relation is encoded as genericRelation whereas FRED represents it through subClassOf. As shown in Table 6, both relations essentially correspond to an is-a hierarchical relation.

Text 4 achieves the highest relation F1-score (5.56%) and the highest accuracy (34.62%). Although nine matching relations are reported, only four represent unique matches. FRED represents the relation between vaginal delivery and vaginal through the property hasQuality, thereby treating vaginal as an attribute that characterises the delivery procedure. In

SDL, by contrast, the relation between vaginal delivery and vagina is manually encoded as `spatialRelation`, focusing on the anatomical location involved in the procedure. Similarly, FRED relates urinary catheter to bladder through the predicate `drain`, reflecting the functional role expressed in the source text, whereas SDL assigns a `spatialRelation`. FRED also links caesarean delivery to surgery through `useOf`, while in SDL terminology the same connection is encoded as a `genericRelation`. Both representations capture relevant aspects of the concept.

These matches exemplify the different modelling choices adopted by the two systems. The examples illustrate how FRED tends to encode functional and linguistic-semantic relations, whereas SDL prioritises conceptual relations.

Similar to Text 2, Text 5 and Text 7 also present matching relations where both systems recognise a relation between concepts, but they classify differently according to their respective modelling principles. In Text 5 only one relation is found. FRED relates prostaglandins and inflammation through the predicate `involved`, reflecting participation in a biological process. In SDL, the relations is manually encoded as an `activityRelation`, which provides a more abstract conceptual representation of the biological role of prostaglandins.

```
{'Fred:': [['prostaglandins'], 'involved', ['inflammation']], 'SDL: ': [['prostaglandin', 'PG'], 'activityRelation ', ['inflammation']]}
```

Text 7 also presents only one unique relation out of two matching relations found. FRED represents symptoms as the `mostProminentSymptomOf` major depressive disorder, whereas in SDL the relationship is manually classified as an `activityRelation`.

```
{'Fred:': [['symptoms'], 'mostProminentSymptomOf', ['MDD']], 'SDL: ': [['symptom'], 'activityRelation ', ['major depressive disorder', 'MDD', 'depression']]}
```

In Text 8 recurring pattern emerge across analysed text, whereby `genericRelation` in SDL corresponds closely to `subClassOf` in FRED.

```
{'Fred:': [['idiopathic gastroparesis', 'Idiopathic Gastroparesis'], 'subClassOf', ['gastroparesis']], 'SDL: ': [['idiopathic gastroparesis', 'gastroparesis with no known cause'], 'genericRelation ', ['gastroparesis', 'delayed gastric emptying']]}
```

```
{'Fred:': [['Diabetic Gastroparesis', 'diabetic gastroparesis'], 'subClassOf',
['gastroparesis']], 'SDL: ': [['diabetic gastroparesis'], 'genericRelation ', ['gastroparesis',
'delayed gastric emptying']]}
```

The FRED extraction connects medications and vomiting through the predicate *reduce*, directly reflecting the therapeutic action described in the source text, while in SDL the same connection is represented through a more general *activityRelation*.

```
{'Fred:': [['medications'], 'reduce', ['vomiting']], 'SDL: ': [['medication', 'medicine',
'pharmaceutical drug', 'drug'], 'activityRelation ', ['vomiting', 'emesis']]}
```

Text 9 also contains only one unique matching relation out of two reported matches. The relations between *inflammation* and *chronic inflammation* are analysed differently by the two approaches. FRED establishes a subtype, while in SDL is encoded as a *propertyRelation*, treating chronicity as a characteristic of *inflammation*. Both interpretations are defensible and this example illustrates how the two approaches may conceptualise the same phenomenon at different levels of abstraction.

```
{'Fred:': [['inflammation'], 'type', ['Chronic Inflammation', 'chronic inflammation']],
'SDL: ': [['inflammation'], 'propertyRelation ', ['chronic inflammation']]}
```

Text 10 again illustrates the correspondence between FRED's *subClassOf* relation and SDL's *genericRelation* in both unique matching relationships out of three common relations.

```
{'Fred:': [['Infectious Disease', 'infectious disease'], 'subClassOf', ['Disease', 'disease']],
'SDL: ': [['infectious disease', 'transmissible disease', 'communicable disease'],
'genericRelation ', ['illness', 'disease']]}
```

```
{'Fred:': [['Communicable Disease', 'communicable disease'], 'subClassOf', ['Disease',
'disease']], 'SDL: ': [['infectious disease', 'transmissible disease', 'communicable
disease'], 'genericRelation ', ['illness', 'disease']]}
```

The lowest performance regarding the term extraction was observed in Text 6, in fact no matching relations between FRED and SDL were identified. As shown in Table 8, the Jaccard Index, precision and F1-score are all equal to 0%. However, as shown in Table 7, recall for term extraction reaches 51.35% and F1-score is 18.63%.

This result suggests that FRED was able to identify a substantial proportion of the SDL concepts while also generating a large number of additional concepts not considered relevant in the SDL analysis. However, the comparatively low precision scores indicate that many extracted terms did not correspond to the manually analysed SDL terms. This discrepancy reflects the broader extraction scope of FRED and the significant quantitative difference between the two datasets. Consequently, the overlap between the two datasets remained moderate despite the relatively high recall scores.

Several examples illustrate this divergence. In the first example, FRED links Development to the DBpedia entity Prenatal development through the relation `<owl#equivalentClass>`.

```
<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#Development>
    a      <http://www.w3.org/2002/07/owl#Class> ;
<http://www.ontologydesignpatterns.org/ont/fred/pos.owl#boxerpos>
<http://www.ontologydesignpatterns.org/ont/fred/pos.owl#n>
> ;
    <http://www.w3.org/2002/07/owl#equivalentClass>
<http://dbpedia.org/resource/Prenatal_development> .
```

Similar cases are represented by the relation `<rdf-schema#subClassOf>` between concepts BirthControlMethod and ControlMethod and also OestrogenReceptorModulator and OestrogenReceptor.

```
<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#BirthControlMethod>

    <http://www.w3.org/2000/01/rdf-schema#subClassOf>
<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#ControlMethod> .
```

The event structures generated by FRED provide another illustration of its linguistic orientation. The predicates, such as suppress, are modelled through semantic roles, assigning method as Agent and ovulation as Patient relations. From the perspective of semantic role labelling, this representation is accurate because it captures the proposition that a contraceptive method suppresses ovulation. This suggests that FRED occasionally transforms textual expressions into ontological entities without sufficient conceptual normalisation.

```

<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#OestrogenReceptorModulator>
    a      <http://www.w3.org/2002/07/owl#Class> ;
    <http://www.w3.org/2000/01/rdf-schema#subClassOf>
<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#ReceptorModulator> ;

<http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#associatedWith>
<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#OestrogenReceptor> .

<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#HormonalContraceptionRefer>
    <http://www.w3.org/2000/01/rdf-schema#subClassOf>
<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#ContraceptionRefer> .

<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#suppress_1>
    a
<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#Suppress> ;

<http://www.ontologydesignpatterns.org/ont/vn/abox/role/Agent>
<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#method_1> ;

<http://www.ontologydesignpatterns.org/ont/vn/abox/role/Patient>
<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#ovulation_1> .

```

A similar tendency can be observed in relations based on `hasQuality` and quantification. While these relations preserve information present in the source text, they generally represent

descriptive linguistic properties rather than conceptual relations relevant to terminology management.

```
<http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#hasQuality>
<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#Cervical> ;
<http://www.ontologydesignpatterns.org/ont/fred/quantifiers.owl#hasQuantifier>
<http://www.ontologydesignpatterns.org/ont/fred/quantifiers.owl#multiple> .
```

An additional example concerns the `<owl#sameAs>` relation between synthetic and chemical synthesis. From an ontological perspective, this mapping is problematic because the adjective synthetic and the process chemical synthesis do not denote the same concept. The former is an adjective describing the artificial production of an entity, while the latter refers to a process through which chemical compounds are created. The relation therefore appears to be based on lexical similarity rather than true conceptual equivalence. This example suggests that the entity-linking component occasionally overgeneralises semantic correspondences when linking concepts to external knowledge bases such as DBpedia.

```
<http://www.ontologydesignpatterns.org/ont/fred/domain.owl#Synthetic>
<http://www.ontologydesignpatterns.org/ont/fred/pos.owl#boxerpos>
<http://www.ontologydesignpatterns.org/ont/fred/pos.owl#> ;
    <http://www.w3.org/2002/07/owl#sameAs>
    <http://dbpedia.org/resource/Chemical_synthesis> .
```

The above relation exemplifies how low relation correspondence does not necessarily indicate failure of the system but rather reflects differences in modelling principles and occasional errors in automatic semantic linking. Here, FRED is not extracting a relation from the text itself; it is making an ontological commitment (`owl:sameAs`) that a terminologist would almost certainly reject.

## 6. Discussion

Medicine has been extensively analysed from a terminological perspective in the National Library of Medicine's Unified Language System (UMLS) (National Library of Medicine (US) 2009), which presents an interoperable ontological terminology for biomedicine. The objective of the present study, however, was not to evaluate FRED as a semantic analysis tool per se, but rather to investigate its potential as a fully automatic term and relation extraction tool from natural language texts.

The comparison between the FRED output and the manually created terminology in SDL proved challenging due to the inherent differences between ontologies and terminologies. Ontologies aim to formalise knowledge through machine-interpretable semantic structures, whereas terminologies focus on the conceptual organisation of specialised knowledge for human communication and knowledge management. Consequently, the comparison involved two fundamentally different types of resources; in fact, a manually created terminology represented in TBX format and an automatically generated ontological representation encoded in RDF/Turtle. In the first instance, the comparison was intended to be conducted between two graphical representations; however, the visualization of the output graph by FRED proved too large and complex to analyse systematically.

Across all ten texts, FRED identified a meaningful proportion of the concepts manually identified and represented in SDL. However, the correspondence between extracted and manually encoded relations is limited. The quantitative findings indicate that FRED demonstrates a moderate capacity for concept extraction but a considerably more limited ability to reproduce the conceptual relations identified through manual terminology analysis. While the average F1-score for term extraction reached 28.92%, relation extraction achieved an average F1-score of only 2.15% (Table 9). These findings suggest that FRED is more successful in identifying relevant domain concepts than in reconstructing the conceptual structures produced through terminology work.

Table 9: Summary of Extraction Performance

| Metric           | Terms           | Relations       |
|------------------|-----------------|-----------------|
| Average F1-Score | 28.92%          | 2.15%           |
| Best F1-Score    | 36.88% (Text 8) | 5.56% (Text 4)  |
| Worst F1-Score   | 18.63% (Text 6) | 0.00 % (Text 6) |

These results demonstrate that concept identification is more successful than relation extraction. This supports the hypothesis that FRED may prove useful as a term extraction tool, but not as a terminology analysis substitute.

The findings suggest that the limited correspondence between FRED and SDL is primarily attributable to their distinct epistemological and representational foundations. Although both approaches frequently identify the same concepts, they often conceptualise the relationships between those concepts differently. FRED extracts semantic relations grounded in the linguistic structure of the source text and primarily derives relations from the linguistic and semantic structure of the source text, by preserving predicates, modifiers, and ontological classifications, thereby modelling the semantic structure of language through RDF/OWL formalisation. The terminology analysis represented in SDL, by contrast, seeks to establish conceptual relations that reflect domain knowledge and support the construction of coherent concept systems.

Consequently, the low F1-scores observed in the quantitative analysis may be interpreted not merely as indicators of extraction performance, but also as evidence of the differing modelling principles underlying the two approaches. The complete absence of overlap observed in Text 6 further illustrates the different representational principles underlying ontology-driven semantic extraction and terminology-driven conceptual modelling.

The analysis also highlights one of FRED's principal strengths, namely its ability to establish semantic links that extend beyond the explicit lexical content of the source text. Several examples demonstrate that FRED successfully links concepts to external knowledge resources such as DBpedia and identifies semantic correspondences between concepts that do not occur verbatim in the analysed texts. This capability may be particularly valuable in biomedical domains, where synonymy and terminological variation are common.

At the same time, the results reveal several limitations of automatic ontology generation. Some relations generated by FRED appear to be driven primarily by linguistic surface structures rather than by domain-specific conceptual knowledge. In other cases, entity-linking operations produce questionable equivalences between concepts that are semantically related but not identical. Furthermore, the parsing process frequently generates duplicate relations. While such repetitions may be acceptable in ontology construction, they are generally not represented in terminology work, where concept systems aim to provide a concise and non-redundant representation of domain knowledge. These observations emphasise the continued importance

of expert validation and manual review when using automatically generated semantic resources. The occurrence of duplicate relations also provides insight into the functioning of FRED, illustrating its extraction process and further underlining the necessity of qualitative human evaluation.

While FRED successfully generates a large number of semantic relations, only a limited subset corresponds to the conceptual relations manually encoded in the terminology represented in SDL. The low F1-scores, precision and recall values demonstrate that most extracted relations were not considered relevant during the manual analysis or did not correspond to the relation types employed in the manual terminology work established in SDL.

The findings are consistent with previous research highlighting the conceptual differences between ontologies and terminologies. As noted by Seppälä (2012), the nature of conceptual characteristics is not always formally specified, leaving terminologists responsible for defining and structuring concept systems. This flexibility contrasts with the formal semantic commitments required by ontology-based systems, which often prioritise logical consistency and machine readability over terminological usability.

As illustrated in Section 5.2, FRED can establish semantic links between synonymous concepts by linking them to external resources even though a term does not occur explicitly in the source text. This behaviour illustrates one of the principal advantages of ontology-based semantic enrichment.

Several methodological limitations should also be acknowledged. The comparison was initially intended to be conducted through graphical representations of the extracted knowledge structures. However, the visualisations generated by FRED proved too large and complex for systematic analysis, making a direct graph-based comparison impractical. Consequently, the comparison was performed on the underlying structured data instead. In addition, technical issues were encountered during the conversion of terminology resources. The original plan was to convert the XML files generated by MemoQ using the MultiTerm XML to TBX Converter; however, the application was unavailable during the study. As a result, the term bases were converted to TBX format using SDL MultiTerm.

FRED generates ontology-oriented relations such as `equivalentClass`, `subClassOf`, `associatedWith`, `hasQuality` and `Agent/Patient` relations. Although the correspondence between FRED and SDL is lower than initially expected, the results indicate that FRED is capable of

identifying a number of conceptually meaningful relationships. The results support three principal conclusions. First, FRED proved moderately effective for concept extraction, whereas relation extraction remained substantially more limited. Second, FRED is considerably less effective at reproducing the conceptual relations identified through the manual terminology analysis. This is reflected in the average F1-Score 2.15% and in the fact that six texts out of ten contain only one unique matching relation, while one text presents no matching relation at all. Third, the limited correspondence in relation extraction should not be intended as an extraction failure, but rather as evidence of the different epistemological foundations of ontology engineering and terminology work.

In many cases, FRED encodes linguistic semantics by preserving the syntactic structure of the source text, whereas the manually created terminology represented in SDL encodes conceptual semantics. Across the analysed texts, recurring patterns emerge. For example, FRED relation `subClassOf` corresponds to the `genericRelation` identified in SDL terminology, `hasQuality` corresponds partially to `propertyRelation`, and `cause` corresponds to `causalRelation`. At the same time, FRED frequently encodes linguistic-semantic relations, such as `reduce`, `drain`, `made`, `involved`, whereas in SDL, conceptual relations are manually encoded for terminology management.

Overall, the study demonstrates that FRED can serve as a useful support tool for terminology work, particularly during the initial stages of concept identification and semantic enrichment. However, the automatic extraction of conceptual relations remains insufficient to replace manual terminology analysis. The findings suggest that the most effective approach may be a hybrid workflow in which ontology-based extraction tools support, rather than replace, expert terminological analysis.

## 7. Conclusion

The aim of this master's thesis was to investigate the potential of FRED as an automatic term and relation extraction tool for terminology work in the biomedical domain. To achieve this purpose, concepts and relations extracted automatically by FRED from biomedical texts were compared with terminological resources manually developed in SDL Multiterm. The analysis was conducted first through a quantitative analysis and then through qualitative analysis of matching concepts and conceptual relations.

The figures demonstrate that FRED is capable of identifying a large number of relevant domain concepts and establishing semantic connections between them. However, the comparison between the two sources reveals a moderate level of correspondence in term extraction, reaching an average F1-score of 28.92%, while relation extraction achieved a lower average F1-score of 2.15%. FRED is therefore more effective in identifying relevant concepts than in reproducing conceptual structures relevant to terminology work.

A closer qualitative investigation revealed that the low matching results obtained through Python parsing do not primarily originate from extraction errors, but rather in inherent different representational principles underlying the two approaches. FRED extracts semantic relations from the linguistic and semantic structure of the source text and formalises them through ontology-oriented relations such as `subClassOf`, `equivalentClass`, and semantic role relations. Terminology work, on the other hand, aims to construct concept systems, in this case by exploiting the functions provided by SDL Multiterm. Therefore, the two approaches often identify similar concepts while conceptualising their relationships in different ways.

The study also revealed both advantages and disadvantages. One of the main advantages is FRED's ability to identify semantic correspondences and to establish links to external knowledge resources, even when concepts were not explicitly present in the analysed texts. Such functionality may be useful in specialised domains characterised by extensive synonymy and terminological variation. Secondly, the high recall values observed throughout the analysis suggest that FRED can extract relevant domain concepts.

On the other hand, the comparison also shows disadvantages, which stem from the inherent structural differences between ontologies and terminologies, thereby limiting the degree of direct correspondence between the two resources. First, FRED generates a large number of relations that reflect linguistic and syntactic structures rather than conceptual

relations relevant to terminology work. Second, the concept relation extraction process frequently produces duplicate relations, highlighting the necessity of manual review.

In conclusion, the observation suggests that FRED can provide valuable support during the preliminary stages of terminology work, particularly for concept identification and semantic enrichment. However, automatic ontology extraction remains insufficient to replace manual terminology analysis, when the objective is the building of concept systems. Rather than a substitute for expert terminology work, FRED should be exploited as a complementary tool that can assist terminologists in identifying concepts and potential semantic relationships. Future research could explore hybrid workflows that combine automatic ontology extraction and human analysis, as well as integrate methods for improving the alignment between ontology-based semantic relations and terminological concept systems.

## 8. Bibliography

- Agustini, Alexandre; Gamallo, Pablo & Lopes, Gabriel P. (2001). Selection Restrictions Acquisition for Parsing Improvement. In: *International Conference on Applications of Prolog*, Springer, 129–143.
- Angele, Jürgen; Kiffer, Michael; Lausen, Georg (2009). Ontologies in F-Logic. In: *Handbook on Ontologies*, 2nd Edition, Springer, Heidelberg, 45-70.
- Antia, Bassey Edem (2000). *Terminology and Language Planning: An Alternative Framework of Practice and Discourse*. Amsterdam/Philadelphia: John Benjamins.
- Arntz, Reiner; Picht, Heribert & Mayer, Felix (2009). *Einführung in die Terminologearbeit*, Hildesheim/Zürich/New York: Georg Olms Verlag.
- Asim, Muhammad Nabeel; Wasim, Muhammad; Khan, Muhammad Usman Ghani; Mahmood, Waqar; Abbasi, Hafiza Mahnoor (2018). A Survey of Ontology Learning Techniques and Applications. *Database*, 2018.
- Atapattu, Thushari; Falkner, Katrina & Falkner, Nickolas (2017). A Comprehensive Text Analysis of Lecture Slides to Generate Concept Maps. In: *Computers & Education*, 115, 96–113.
- Aussenac-Gilles, Nicole, Szulman, Sylvie & Depres, Sylvie (2008). The Terminae Method and Platform for Ontology Engineering from Texts. In: Paul Buitelaar and Philipp Cimiano (eds.), *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, Amsterdam: IOS Press, 199-223.
- Baader, Franz; Horrocks, Ian; Sattler, Ulrike (2009). Description Logics. In: *Handbook on Ontologies*, (2nd Edition), Springer-Verlag Berlin/Heidelberg, 21-43.
- Bandrowski, Anita; Brinkman, Ryan; Brochhausen, Mathias; Brush, Matthew H.; Bug, Bill; Chibucos, Marcus C.; Clancy, Kevin; Courtot, Mélanie; Derom, Dirk; Dumontier, Michel; Fan, Liju; Fostel, Jennifer; Fragoso, Gilberto; Gibson, Frank; Gonzalez-Beltran, Alejandra; Haendel, Melissa A.; He, Yongqun; Heiskanen, Mervi; Hernandez-Boussard, Tina; Jensen, Mark; Lin, Yu; Lister, Allyson L.; Lord, Phillip; Malone, James; Manduchi, Elisabetta; McGee, Monnie; Morrison, Norman; Overton, James A.; Parkinson, Helen; Peters, Bjoern; Roca-Serra, Philippe; Ruttenberg, Alan; Sansone, Susanna-Assunta; Scheuermann, Richard H.; Schober, Daniel; Smith, Barry; Soldatova,

- Larisa N.; Stoeckert, Christian J. Jr; Taylor, Chris F.; Torniai, Carlo; Turner, Jessica A.; Vita, Randi; Whetzel, Patricia L. & Zheng, Jie (2016). The Ontology for Biomedical Investigations. *PLoS ONE*, 11(4), e0154556.
- Barsalou, Lawrence W. (1992). Frames, Concepts, and Conceptual Fields. In: Adrienne Lehrer and Eva Feder Kittay (eds.), *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*. Lawrence Erlbaum, 21-74.
- Berkhin, Pavel (2006). A survey of clustering data mining techniques. In: *Grouping Multidimensional Data*, Springer, 25–71.
- Bos, Johan (2008). Wide-Coverage Semantic Analysis with Boxer. In: J. Bos and R. Delmonte, (eds.), *Conference on Semantics in Text Processing (STEP)*. Venice: College Publications, 277–286.
- Budin, Gerhard (1996). *Wissensorganisation und Terminologie: Die Komplexität und Dynamik Wissenschaftlicher Informations- und Kommunikationsprozesse*. Narr: Tübingen.
- Buitelaar, Paul; Cimiano, Philipp & Magnini, Bernardo (2005). Ontology learning from text: an overview. In: *Ontology Learning from Text: Methods, Evaluation and Applications*. Amsterdam, IOS Press, 123, 3–12
- Cabré Castellví, Maria Teresa (2003). Theories of terminology: Their description, prescription and explanation. In: *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(2) 163 – 199.
- Ciaramita, Massimiliano; Gangemi, Aldo; Ratsch, Esther; Šarić, Jasmin & Rojas, Isabel (2005). Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In: *IJCAI'05 (Proceedings of the 19<sup>th</sup> international joint conference on Artificial Intelligence*, 659–664.
- Cimiano, Philipp; Mädche, Alexander; Staab, Steffen & Völker, Johanna (2009). Ontology Learning. In: *Handbook on Ontologies*, 2nd ed. Springer, 245-267.
- Cunningham, Hamish; Gaizauskas, Robert; Humphreys, Kevin & Wilks, Yorick (1999). Experience with a Language Engineering Architecture: Three Years of GATE. In: *Proceedings of the AISB'99 Workshop on Reference Architectures and Data Standards for NLP*.

- Drewer, Petra; Massion, François & Pulitano, Donatella (2017). Was haben Wissensmodellierung, Wissensstrukturierung, künstliche Intelligenz und Terminologie miteinander zu tun?. In: *Deutscher Terminologie-Tag (DTT): DDT-Symposion 2017. Terminologie und KI im digitalen Wandel*. 1–28.
- Drouin, Patrick; Grabar, Natalia; Hamon, Thierry; Kageura, Kyo & Takeuchi, Koichi (2018). Computational terminology and filtering of terminological information. In: *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 24 (1), 1-6.
- De Bessé, Bruno (1997). Concept descriptions. In: Budin, Gerhard & Wright, Sue Ellen (eds.) *Handbook of Terminology Management*. Amsterdam: John Benjamins Publishing, Volume 1, 63-87.
- Dhillon, Inderjit S.; Mallela, Subramanyam & Kumar, Rahul (2003). A divisive information-theoretic feature clustering algorithm for text classification. In: *Journal of Machine Learning Research* 3, 1265–1287.
- Drymonas, Efthymios G. (2009). *Ontology learning from text based on multiword term concepts: the OntoGain method*. Master Thesis, Technical University of Crete, Greece.
- Dury, Pascaline (2005). Terminology and specialised translation: The relevance of the diachronic approach. In: *LSP and Professional Communication* 5(1), 31-42.
- Faber, Pamela (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/New York: Mouton de Gruyter.
- Fielding, Roy Thomas (2000). *Architectural Styles and the Design of Network-based Software Architectures*. PhD Thesis. University of California, Irvine. Available on <https://ics.uci.edu/~fielding/pubs/dissertation/top.htm> (accessed on 20.05.2020)
- Fillmore, Charles J. (1976). Frame semantics and the nature of language. In: *Annals New York Academy of Sciences*, 208, 20-32.
- Felber, Helmut (1984). *Terminology Manual*. Paris: Infoterm/UNESCO.
- Fortuna, Blaž, Lavrac, Nada & Velardi, Paola (2008). Advancing topic ontology learning through term extraction. In: *PRICAI 2008 Proceedings*. Springer, 626–635.

- Frantzi, Katerina T, Ananiadou, Sophia & Mima, Hideki (2000). Automatic recognition of multiword terms: The C-value/NC-value method. In: *Research and Advanced Technology for Digital Libraries: Second European Conference, ECDL'98 Heraklion, Crete, Greece September 21–23, 1998 Proceedings*, 115–130.
- Fridsma, Douglas B.; Evans, Julie; Hastak, Smita & Mead, Charles N. (2008). The BRIDG Project: A Technical Report. In: *Journal of the American Medical Informatics Association*, 15(2), 130–137.
- Fung, Kin Wah & Bodenreider, Olivier (2023). Knowledge representation and ontologies. In: *Clinical research informatics*. Springer, 367-388
- Gamallo, Pablo, Agustini, Alexandre & Lopes, Gabriel P. (2003). Learning subcategorization information to model a grammar with “co-restrictions”. In: *Modélisation probabiliste du langage naturel. TAL: Traitement automatique des langues*, 44, 93–117.
- Gangemi, Aldo (2005). Ontology design patterns for semantic web content. In: Y. Gil et al., (eds.), *Proceedings of the 4th Int. Semantic Web Conference, ISWC 2005, volume 3729 of LNCS*, Springer. 262–276.
- Gangemi, Aldo (2013). A Comparison of Knowledge Extraction Tools for the Semantic Web. In: *ESWC 2013: The Semantic Web: Semantics and Big Data*, 351-366.
- Gangemi, Aldo; Draicchio, Francesco; Presutti, Valentina, Nuzzolese, Andrea Giovanni & Reforgiato, Diego (2013a). A Machine Reader for the Semantic Web. In: *Proceedings of the ISWC 2013 Posters & Demonstrations Track within the 12th International Semantic Web Conference (ISWC 2013)*, 1035, 149-152.
- Gangemi, Aldo; Hassan, Ehab; Presutti, Valentina & Reforgiato Recupero, Diego (2013b). FRED as an Event Extraction Tool. In: *Proceedings of the 3rd Workshop on Detection, Representation and Exploitation of Events in the Semantic Web (DeRiVE 2013)*. 1123, 14–17.
- Gangemi, Aldo; Presutti, Valentina; Reforgiato Recupero, Diego; Nuzzolese, Andrea Giovanni; Draicchio, Francesco & Mongiovì, Misael (2017). Semantic Web Machine Reading with FRED. In: *Semantic Web Journal*, 8(6), 873-893.
- Genesereth, Michael R. & Nilsson, Nils J. (1987). *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann, Los Altos, CA, 1987.

- Gómez-Pérez, Asunción & Manzano-Macho, David (2004). An overview of methods and tools for ontology learning from texts. In: *The Knowledge Engineering Review*, 19(3), 187–212.
- Gromann, Dagmar (2013). Terminology meets the multilingual Semantic Web: A semiotic comparison of ontologies and terminologies. In: *Languages for Special Purposes in a Multilingual, Transcultural World Proceedings of the 19th European Symposium on Languages for Special Purposes*, 8-10 July 2013, Vienna, Austria, 417-418.
- Gromann, Dagmar (2014). A Model and Method to Terminologize Existing Domain Ontologies. In: *Proceedings of the 11th International Conference on Terminology and Knowledge Engineering (TKE 2014), "Ontology, Terminology and Text Mining", Berlin, Germany, 19–21 June 2014*. Lyon: Centre pour la Communication Scientifique Directe (CCSD), 127–136.
- Gromann, Dagmar (2015). *Design and evaluation of an ontology-terminology model for the interchange of terminological data: an approach to multilingual conceptual modelling as exemplified in the domain of finance*. PhD thesis, University of Vienna.
- Gromann, Dagmar (2020). Ontologien als terminologische Begriffssysteme. In: Drewer, P., Mayer, F. & Pulitano, D. (eds): *Terminologie: Industrie, Information, Intelligenz*. Deutscher Terminologie-Tag e.V. München, Karlsruhe, Bern. 39-48
- Guarino, Nicola; Oberle, Daniel; Staab, Steffen (2009). What is an ontology?. In: *Handbook on Ontologies*, 2nd Edition, Springer, Berlin/Heidelberg, 1-17.
- Hearst, Marti A. (1992). Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of the 14th International Conference on Computational Linguistics. Nantes, France, 1992, 23-28 August 1992*, 539-545.
- Hearst, Marti A. (1998). Automated discovery of WordNet relations, In: Fellbaum, Christiane (ed.) *WordNet: an electronic lexical database*, MIT Press, 131–153.
- Hwang, Chung Hee (1999). Incompletely and imprecisely speaking: using dynamic ontologies for representing and retrieving information. In: *KRDB, CEUR-WS, 6<sup>th</sup> International Workshop on Knowledge Representation meets Database (KRDB '99)*, 14–20.
- Hellmann, Sebastian; Lehmann, Jens; Auer, Sören & Brümmer, Martin (2013). Integrating NLP using Linked Data. In: H. Alani, L. Kagal, A. Fokoue, P. T. Groth, C. Biemann, J. X.

- Parreira, L. Aroyo, N. F. Noy, C. Welty & K. Janowicz (eds.) *The Semantic Web – ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II, volume 8219 of Lecture Notes in Computer Science*. Springer, 98–113.
- Hitzler, Pascal; Krötzsch, Markus & Rudolph, Sebastian (2009). *Foundations of Semantic Web Technologies*. CRC.
- Ismail, Rohana; Bakar, Zainab Abu & Rahman, Nurazzah Abdul (2015). Extracting knowledge from English translated Quran using NLP pattern. In: *Jurnal Teknologi*, 77, 67–73.
- Ismail, Rohana; Bakar, Zainab Abu & Rahman, Nurazzah Abdul (2017). An ontology learning framework for Quran. In: *Advanced Science Letters*, 23, 4175–4178.
- ISO 704: 2009. *Terminology Work - Principles and Methods*, International Organization for Standardization. Geneva: International Organization for Standardization.
- ISO 1087: 2019. *Terminology work and terminology science — Vocabulary*, International Organization for Standardization. Geneva: International Organization for Standardization.
- ISO/TC 37: 1995. *Terminology Work - Principles and Methods*. Geneva: International Organization for Standardization.
- ISO 30042: 2019. *Systems to manage terminology, knowledge and content — TermBase eXchange (TBX)*. Geneva: International Organization for Standardization.
- ISO 5725-1:1994. *Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions*. Geneva: International Organization for Standardization.
- Jaccard, Paul (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société vaudoise des sciences naturelles*, 37, 547–579.
- Janowicz, Krzysztof; Gangemi, Aldo; Hitzler, Pascal; Krisnadhi, Adila; Presutti, Valentina (2016). Ontology Design Patterns in a Nutshell. In: Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi, Valentina Presutti (eds.), *Ontology Engineering with Ontology Design Patterns: Foundations and Applications (Studies on the Semantic Web)*, 25. Amsterdam: IOS Press, 1–12.

- Jiang, Xing & Tan, Ah-Hwee (2010). CRCTOL: a semantic-based domain ontology learning system. In: *Journal of the American Society for Information Science and Technology*, 61(1), 150–168.
- Kamp, Hans (1981). A theory of truth and semantic representation. In: J. A. G. Groenendijk, T. M. V. Janssen, and M. B. J. Stokhof, (eds.), *Formal Methods in the Study of Language*, 277–322.
- Kang, Sung Ku; Patil, Lalit; Rangarajan, Arvind; Moitra, Abha; Jia, Tao; Robinson, Dean M. & Dutta, Debasish (2015). Extraction of manufacturing rules from unstructured text using a semantic framework. In: *ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*.
- Kaushik, Neha & Chatterjee, Niladri (2018). Automatic relationship extraction from agricultural text for ontology construction. In: *Information Processing in Agriculture*, 5(1), 60–73.
- Kless, Daniel; Milton, Simon (2010). Comparison of Thesauri and Ontologies from a Semiotic Perspective. In: T. Meyer, M. A. Orgun & K. Taylor (eds.), *Proceedings of the 6th Australian Ontology Workshop (AOW 2010)*, 35–44.
- Lassila, Ora & McGuinness, Deborah L. (2001). *The Role of Frame-Based Representations on the Semantic Web*, Technical Report, Stanford: Knowledge System Laboratory Report KLS-01-02, Stanford University.
- Lima, Rinaldo; Espinasse, Bernard; Oliveira, Hilário; Ferreira, Rafael; Cabral, Luciano; Filho, Dimas; Freitas, Fred & Gadelha, Renê (2013). An inductive logic programming-based approach for ontology population from the web. In: *DEXA: International Conference on Database and Expert Systems Applications, 24<sup>th</sup> International Conference, DEXA 2013, Prague, Czech Republic*, Springer, 319–326.
- LISA (2008). *Systems to manage terminology, knowledge, and content - TermBase eXchange (TBX)*. Tech. rep. Localization Industry Standards Association.
- Mädche, Alexander & Staab, Steffen (2001). Ontology Learning for the Semantic Web. In: *IEEE Intelligent Systems* 16 (2), 72-79.
- Mädche, Alexander & Staab, Steffen (2004). Ontology learning. In: *Handbook on Ontologies* (1st Edition), Springer Berlin/Heidelberg: 173-190.

- Mayer, Felix (1998). Eintragsmodelle für Terminologische Datenbanken - Ein Beitrag zur Übersetzungsorientierten Terminographie. In: Hartwig Kalverkämper (eds.), *Forum für Fachsprachenforschung*, 44, Tübingen: Narr.
- Mora, Sara, Roberta Gazzarata, Bernd Blobel, Ylenia Murgia & Mauro Giacomini (2024). Transforming Ontology Web Language Elements into Common Terminology Service 2 Terminology Resources. In: *Journal of Personalized Medicine* 14(7), 676.
- National Library of Medicine (US) (2009). *UMLS® Reference Manual*. Bethesda.
- Navigli, Roberto & Velardi, Paola (2002). Semantic interpretation of terminological strings. In: *Proceedings of the Sixth International Conference on Terminology and Knowledge Engineering*, 95–100.
- Navigli, Roberto; Velardi, Paola & Gangemi, Aldo (2003). Ontology learning and its application to automated terminology translation. In: *IEEE Intelligent Systems*. 18, 22–31.
- Njike-Fotzo, Hermine & Gallinari, Patrick (2004). Learning generalization/specialization relations between concepts–application for automatically building thematic document hierarchies. In: *Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, 143–155.
- Noy, Natalya F.; Shah, Nigam H.; Whetzel, Patricia L.; Dai, Benjamin; Dorf, Michael; Griffith, Nicholas; Jonquet, Clement; Rubin, Daniel L.; Storey, Margaret-Anne; Chute, Christopher G. & Musen, Mark A. (2009). BioPortal: Ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(Web Server issue), W170–W173
- Nuopponen, Anita (1994). *Begreppssystem för terminologisk analys [Concept systems for terminological analysis]*. Vaasa: University of Vaasa, Acta Wasaensia.
- Nuopponen, Anita (2010). Methods of concept analysis – towards systematic concept analysis (part 2 of 3). In: *The LSP Journal – Language for Special Purposes, Professional Communication, Knowledge Management and Cognition 1(1)*. 5–14.
- Nuopponen, Anita (2018). Terminological Concept Systems. In: John Humbley, Gerhard Budin, Christer Laurén (eds.), *Languages for Special Purposes: An International Handbook*. Berlin/Boston: De Gruyter, 453–468.

- Ogden, Charles Kay & Richards, Ivor Armstrong (1923). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Cambridge University Press.
- Panchenko, Alexander; Faralli, Stefano; Ruppert, Eugen; Remus, Steffen; Naets, Hubert; Fairon, Cédric; Ponzetto, Simone Paolo; Biemann, Chris (2016). Taxi at semeval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In: *Proceedings of the 10<sup>th</sup> International Workshop on Semantic Evaluation (SemEval-2016)*, 1320–1327.
- Peroni, Silvio & Vitali, Fabio (2009). Annotations with EARMARK for arbitrary, overlapping and out-of-order markup. In: Borghoff, Uwe M. & Chidlovskii, Boris (eds.) *Proceedings of the 2009 ACM Symposium on Document Engineering*, 2009, 171–180.
- Powers, David M. W. (2007). *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. Technical Report SIE-07-001, Adelaide: Flinders University. Available online: <http://fac.flinders.edu.au/items/90ac6613-2b25-4c9a-887a-7194da37e79c>
- Pozzi, María (2001). The Terminological Definition: Conflicts Between Theory and Practice. In: Mayer, Felix (ed.), *Language for Special Purposes: Perspectives for the New Millennium: Linguistics and Cognitive Aspects, Knowledge Representation and Computational Linguistics, Terminology, Lexicography and Didactics*; Tübingen: Narr, 272-281.
- Prampolini, Massimo (2013). *Ferdinand de Saussure*. Carocci, Roma.
- Presutti, Valentina; Draicchio, Francesco & Gangemi, Aldo (2012). Knowledge Extraction Based on Discourse Representation Theory and Linguistic Frames. In: *Proceedings of Knowledge Engineering and Knowledge Management, 18<sup>th</sup> International Conference (EKAW 2012)*, Galway City, Ireland, Springer, 114-129.
- Saussure, Ferdinand de (2011). *Course in General Linguistics*. Translated by Wade Baskin. Columbia University Press, 65-101.
- Sen, Sagnika; Tao, Jie & Deokar, Amit V. (2015). On the role of ontologies in information extraction. In: *Reshaping Society through Analytics, Collaboration, and Decision Support*, Springer, 115–133.

- Sim, Ida; Tu, W Samson; Carini, Simona; Lehmann, Harold P.; Pollock, Brad H.; Peleg, Mor & Wittkowski, Knut M. (2014). The Ontology of Clinical Research (OCRe): An informatics foundation for the science of clinical research. In: *Journal of Biomedical Informatics*, 52, 78–91.
- Smith, Barry, Ceusters, Werner, & Temmerman, Rita (2005). Wüsteria. In: *Studies in Health Technology and Informatics*, 116, 647-52.
- Smith, Barry, Kusnierczyk, Wacław, Schober, Daniel & Ceusters, Werner (2006). Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain. In: Bodenreider, Oliver (ed.), *KR-MED 2006 Proceedings*, Baltimore, Maryland, November 8<sup>th</sup> 2006; Knowledge Representation of the American Medical Informatics Association (AMIA), 57-66.
- Snow, Rion; Jurafsky, Daniel & Ng, Andrew Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In: *NIPS Proceedings 2005, Advances in Neural Information Processing Systems 18*, 1297–1304.
- Sombatsrisomboon, Ratanachai; Matsuo, Yutaka & Ishizuka, Mitsuru (2003). Acquisition of hypernyms and hyponyms from the WWW. In *Proceedings of the 2nd International Workshop on Active Mining (AM-2003)*.
- Sordo, Mohamed; Oramas, Sergio; Espinosa-Anke, Luis (2015). Extracting relations from unstructured text sources for music recommendation. In: Biemann, Chris; Handschuh, Siegfried; Freitas, André; Mezziane, Farid & Métais, Elisabeth (eds.), *20th International Conference on Applications of Natural Language to Information Systems (NLDB 2015)*. Cham: Springer, 369-382
- Steedman, Mark (2000). *The Syntactic Process*. MIT Press, Cambridge, MA, USA.
- Suonuuti, Heidi (1997). *Guide to Terminology*. Nordterm
- Temmerman, Rita (1998). Why traditional terminology impedes a realistic description of categories and terms in the life sciences. In: *Terminology* 5(1), January 1998, 77–92.
- Temmerman, Rita (2000). *Towards new ways of terminology description: the sociocognitive approach*. Amsterdam/ Philadelphia, PA: J. Benjamins.
- Tolkien, J. R. R. (1927). Philology: General Works. In: *The Year's Work in English Studies*, 6.

- Uschold, Michael & Grüninger, Michael. (2004). Ontologies and Semantics for Seamless Connectivity. In: *SIGMOD Record, Volume 33*, 58-64
- Wright, Sue Ellen (1997). Concept systems. In: Budin, Gerhard & Wright, Sue Ellen (eds.) *Handbook of Terminology Management*. Amsterdam: John Benjamins Publishing, 89-143.
- Wüster, Eugen (1974). Die allgemeine Terminologielehre – ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften. In: *Linguistics, Volume 199, January 1*, 61–106.
- Wüster, Eugen (1975). *Der Unterschied zwischen den logischen und den ontologischen (z. B. partitiven) Beziehungen*. Unpublished. W 639, Wü/Re 1975-02-19, Wüster-Bibliothek, Vienna.
- Wüster, Eugen (1979). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Wien: Springer.
- Zepeda-Mendoza, Marie Lisandra & Resendis-Antonio, Osbaldo (2013). Hierarchical agglomerative clustering. In: Dubitzky, W.; Wolkenhauer, O.; Cho, K.-H. & Yokota, H., *Encyclopedia of Systems Biology*, New York: Springer, 886–887.

## Web sources:

Object Management Group (OMG) (n.d.). Common Terminology Services 2 (CTS2). Available online: <https://www.omg.org/cts2/> (accessed on 25.04.2025).

PlanetCalc (n.d.). Jaccard / Tanimoto coefficient. Available online: <https://planetcalc.com/1664/> (accessed on 22.09.2022).

TermBase eXchange – Tools (n.d.). Available online: <https://www.tbxinfo.net/tools/> (accessed on 08.08.2022).

Wikimedia Commons contributors (2014). Semantic Web Stack. Available online: [https://commons.wikimedia.org/wiki/File:Semantic\\_web\\_stack.svg](https://commons.wikimedia.org/wiki/File:Semantic_web_stack.svg) (accessed on 25.03.2020).

Wikipedia contributors (n.d.). Extraocular muscles. Available online: [https://en.wikipedia.org/wiki/Extraocular\\_muscles](https://en.wikipedia.org/wiki/Extraocular_muscles) (accessed on 08.05.2022).

World Wide Web Consortium (W3C) (2004). Architecture of the World Wide Web, Volume One. W3C Recommendation 15 December 2004. Ian Jacobs, Norman Walsh. Available online: <https://www.w3.org/TR/webarch/> (accessed on 25.05.2020).

World Wide Web Consortium (W3C) (2004). OWL Web Ontology Language Reference. W3C Recommendation 10 February 2004. Available online: <https://www.w3.org/TR/2004/REC-owl-ref-20040210/> (accessed on 31.08.2021).

World Wide Web Consortium (W3C) (2012). OWL 2 Web Ontology Language Primer (Second Edition). W3C Recommendation 11 December 2012. Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Sebastian Rudolph. Available online: <https://www.w3.org/TR/owl2-primer/> (accessed on 03.05.2020).

World Wide Web Consortium (W3C) (2014a). RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation 25 February 2014. Richard Cyganiak, David Wood, Markus Lanthaler. Available online: <https://www.w3.org/TR/rdf11-concepts/> (accessed on 14.02.2020).

World Wide Web Consortium (W3C) (2014b). Turtle – Terse RDF Triple Language. W3C Recommendation 25 February 2014. David Beckett, Tim Berners-Lee, Eric Prud'hommeaux, Gavin Carothers. Available online: <https://www.w3.org/TR/turtle/> (accessed on 14.02.2020).

World Wide Web Consortium (W3C) (n.d.). About W3C. Available online: <https://www.w3.org/Consortium/> (accessed on 25.05.2020).

World Wide Web Consortium (W3C) (n.d.). Mission. Available online:

<https://www.w3.org/Consortium/mission> (accessed on 25.05.2020).

World Wide Web Consortium (W3C) (n.d.). Semantic Web Books. Available online:

<https://www.w3.org/wiki/SwBooks> (accessed on 17.02.2020).

World Wide Web Consortium (W3C) (n.d.). Semantic Web Ontologies. Available online:

<https://www.w3.org/standards/semanticweb/ontology> (accessed on 25.05.2020).

World Wide Web Consortium (W3C) (n.d.). Semantic Web Standards. Available online:

<https://www.w3.org/standards/semanticweb/> (accessed on 25.05.2020).

World Wide Web Consortium (W3C) (n.d.). Standards. Available online:

<https://www.w3.org/standards/> (accessed on 25.05.2020).

## 9. Appendix

### 9.1 Text Sources

#### Text 1: Diabetes Mellitus

Diabetes mellitus (DM), commonly known as diabetes, is a group of metabolic disorders characterized by a high blood sugar level over a prolonged period of time. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. If left untreated, diabetes can cause many complications. Acute complications can include diabetic ketoacidosis, hyperosmolar hyperglycemic state, or death. Serious long-term complications include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, damage to the nerves, and damage to the eyes.

Diabetes is due to either the pancreas not producing enough insulin, or the cells of the body not responding properly to the insulin produced. There are three main types of diabetes mellitus:

Type 1 diabetes results from the pancreas's failure to produce enough insulin due to loss of beta cells. This form was previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes". The loss of beta cells is caused by an autoimmune response. The cause of this autoimmune response is unknown.

Type 2 diabetes begins with insulin resistance, a condition in which cells fail to respond to insulin properly. As the disease progresses, a lack of insulin may also develop. This form was previously referred to as "non insulin-dependent diabetes mellitus" (NIDDM) or "adult-onset diabetes". The most common cause is a combination of excessive body weight and insufficient exercise.

Gestational diabetes is the third main form, and occurs when pregnant women without a previous history of diabetes develop high blood sugar levels.

<https://en.wikipedia.org/wiki/Diabetes> (accessed on 06.04.2020)

#### Text 2: Multiple Sclerosis

Multiple sclerosis (MS) is a demyelinating disease in which the insulating covers of nerve cells in the brain and spinal cord are damaged. This damage disrupts the ability of parts of the nervous system to communicate, resulting in a range of signs and symptoms, including physical, mental,

and sometimes psychiatric problems. Specific symptoms can include double vision, blindness in one eye, muscle weakness, trouble with sensation, or coordination. MS takes several forms, with new symptoms either occurring in isolated attacks (relapsing forms) or building up over time (progressive forms). Between attacks, symptoms may disappear completely; however, permanent neurological problems often remain, especially with the advancement of the disease.

While the cause is unclear, the underlying mechanism is thought to be either destruction by the immune system or failure of the myelin-producing cells. Proposed causes for this include genetics and environmental factors such as being triggered by a viral infection. MS is usually diagnosed based on the presenting signs and symptoms and the results of supporting medical tests.

There is no known cure for multiple sclerosis. Treatments attempt to improve function after an attack and prevent new attacks. Medications used to treat MS, while modestly effective, can have side effects and be poorly tolerated.

[https://en.wikipedia.org/wiki/Multiple\\_sclerosis](https://en.wikipedia.org/wiki/Multiple_sclerosis) (accessed on 06.04.2020)

### **Text 3: Colonoscopy**

Colonoscopy or coloscopy is the endoscopic examination of the large bowel and the distal part of the small bowel with a CCD camera or a fiber optic camera on a flexible tube passed through the anus. It can provide a visual diagnosis (e.g., ulceration, polyps) and grants the opportunity for biopsy or removal of suspected colorectal cancer lesions. Colonoscopy can remove polyps as small as one millimeter or less. Once polyps are removed, they can be studied with the aid of a microscope to determine if they are precancerous or not. [...]

Conditions that call for colonoscopies include gastrointestinal haemorrhage, unexplained changes in bowel habit and suspicion of malignancy. Colonoscopies are often used to diagnose colon cancer, but are also frequently used to diagnose inflammatory bowel disease. In older patients (sometimes even younger ones) an unexplained drop in haematocrit (one sign of anaemia) is an indication that calls for a colonoscopy, usually along with an esophagogastroduodenoscopy (EGD), even if no obvious blood has been seen in the stool (feces).

Fecal occult blood is a quick test which can be done to test for microscopic traces of blood in the stool. A positive test is almost always an indication to do a colonoscopy. In most cases the

positive result is just due to haemorrhoids; however, it can also be due to diverticulosis, inflammatory bowel disease (Crohn's disease, ulcerative colitis), colon cancer, or polyps. Colonic polypectomy has become a routine part of colonoscopy, allowing quick and simple removal of polyps during the procedure, without invasive surgery.

<https://en.wikipedia.org/wiki/Colonoscopy> (accessed on 06.04.2020)

#### **Text 4: Caesarean section**

Caesarean section, also known as C-section, or caesarean delivery, is the use of surgery to deliver babies. A caesarean section is often necessary when a vaginal delivery would put the baby or mother at risk. This may include obstructed labour, twin pregnancy, high blood pressure in the mother, breech birth, or problems with the placenta or umbilical cord. A caesarean delivery may be performed based upon the shape of the mother's pelvis or history of a previous C-section. A trial of vaginal birth after C-section may be possible. The World Health Organization recommends that caesarean section be performed only when medically necessary. Some C-sections are performed without a medical reason, upon request by someone, usually the mother.

A C-section typically takes 45 minutes to an hour. It may be done with a spinal block, where the woman is awake, or under general anaesthesia. A urinary catheter is used to drain the bladder, and the skin of the abdomen is then cleaned with an antiseptic. An incision of about 15 cm (6 inches) is then typically made through the mother's lower abdomen. The uterus is then opened with a second incision and the baby delivered. The incisions are then stitched closed. A woman can typically begin breastfeeding as soon as she is out of the operating room and awake. Often, several days are required in the hospital to recover sufficiently to return home.

[https://en.wikipedia.org/wiki/Caesarean\\_section](https://en.wikipedia.org/wiki/Caesarean_section) (accessed on 06.04.2020)

#### **Text 5: Nonsteroidal anti-inflammatory drugs**

Nonsteroidal anti-inflammatory drugs (NSAIDs) are members of a drug class that reduces pain, decreases fever, prevents blood clots, and in higher doses, decreases inflammation. Side effects depend on the specific drug but largely include an increased risk of gastrointestinal ulcers and bleeds, heart attack, and kidney disease.

The term nonsteroidal distinguishes these drugs from steroids, which while having a similar

eicosanoid-depressing, anti-inflammatory action, have a broad range of other effects. First used in 1960, the term served to distance these medications from steroids, which were particularly stigmatised at the time due to the connotations with anabolic steroid abuse.

NSAIDs work by inhibiting the activity of cyclooxygenase enzymes (COX-1 and/or COX-2). In cells, these enzymes are involved in the synthesis of key biological mediators, namely prostaglandins, which are involved in inflammation, and thromboxanes, which are involved in blood clotting.

The most prominent NSAIDs are aspirin, ibuprofen, and naproxen, all available over the counter (OTC) in most countries. Paracetamol (acetaminophen) is generally not considered an NSAID because it has only minor anti-inflammatory activity. It treats pain mainly by blocking COX-2 and inhibiting endocannabinoid reuptake almost exclusively within the brain, but not much in the rest of the body.

[https://en.wikipedia.org/wiki/Nonsteroidal\\_anti-inflammatory\\_drug](https://en.wikipedia.org/wiki/Nonsteroidal_anti-inflammatory_drug) (accessed on 06.04.2020)

## **Text 6: Hormonal contraception**

Hormonal contraception refers to birth control methods that act on the endocrine system. Almost all methods are composed of steroid hormones, although in India one selective oestrogen receptor modulator is marketed as a contraceptive. The original hormonal method—the combined oral contraceptive pill—was first marketed as a contraceptive in 1960. In the ensuing decades many other delivery methods have been developed, although the oral and injectable methods are by far the most popular. Altogether, 18% of the world's contraceptive users rely on hormonal methods. Hormonal contraception is highly effective: when taken on the prescribed schedule, users of steroid hormone methods experience pregnancy rates of less than 1% per year. Perfect-use pregnancy rates for most hormonal contraceptives are usually around the 0.3% rate or less. Currently available methods can only be used by women; the development of a male hormonal contraceptive is an active research area.

There are two main types of hormonal contraceptive formulations: combined methods which contain both an oestrogen and a progestin, and progestogen-only methods which contain only progesterone or one of its synthetic analogues (progestins). Combined methods work by suppressing ovulation and thickening cervical mucus; while progestogen-only methods reduce the frequency of ovulation, most of them rely more heavily on changes in cervical mucus. The

incidence of certain side effects is different for the different formulations: for example, breakthrough bleeding is much more common with progestogen-only methods. Certain serious complications occasionally caused by oestrogen-containing contraceptives are not believed to be caused by progestogen-only formulations: deep vein thrombosis is one example of this.

[https://en.wikipedia.org/wiki/Hormonal\\_contraception](https://en.wikipedia.org/wiki/Hormonal_contraception) (accessed on 06.04.2020)

### **Text 7: Antidepressants**

Antidepressants are medications used to treat major depressive disorder, some anxiety disorders, some chronic pain conditions, and to help manage some addictions. Common side-effects of antidepressants include dry mouth, weight gain, dizziness, headaches, and sexual dysfunction. Most types of antidepressants are typically safe to take, but may cause increased thoughts of suicide when taken by children, adolescents, and young adults. A discontinuation syndrome can occur after stopping any antidepressant which resembles recurrent depression.

Antidepressants are used to treat major depressive disorder and of other conditions, including some anxiety disorders, some chronic pain conditions, and to help manage some addictions. Antidepressants are often used in combinations with one another. The proponents of the monoamine hypothesis of depression recommend choosing the antidepressant with the mechanism of action impacting the most prominent symptoms—for example, they advocate that people with MDD who are also anxious or irritable should be treated with SSRIs or norepinephrine reuptake inhibitors, and the ones with the loss of energy and enjoyment of life—with norepinephrine and dopamine enhancing drugs.

[...] Almost any medication involved with serotonin regulation has the potential to cause serotonin toxicity (also known as serotonin syndrome) — an excess of serotonin that can induce mania, restlessness, agitation, emotional lability, insomnia and confusion as its primary symptoms.

<https://en.wikipedia.org/wiki/Antidepressant> (accessed on 06.04.2020)

### **Text 8: Gastroparesis**

Gastroparesis (gastro- from Ancient Greek γαστήρ - gaster, "stomach"; and -paresis, πάρεσις -

"partial paralysis"), also called delayed gastric emptying, is a medical disorder consisting of weak muscular contractions (peristalsis) of the stomach, resulting in food and liquid remaining in the stomach for a prolonged period of time. Stomach contents thus exit more slowly into the duodenum of the digestive tract.

Symptoms include nausea, vomiting, abdominal pain, feeling full soon after beginning to eat (early satiety), abdominal bloating, and heartburn. The most common known mechanism is autonomic neuropathy of the nerve which innervates the stomach: the vagus nerve. Uncontrolled diabetes mellitus is a major cause of this nerve damage; other causes include post-infectious and trauma to the vagus nerve.

Diagnosis is via one or more of the following: barium swallow X-ray, barium beefsteak meal, radioisotope gastric-emptying scan, gastric manometry, and esophagogastroduodenoscopy (EGD). Complications include malnutrition, fatigue, weight loss, vitamin deficiencies, intestinal obstruction due to bezoars, and small intestine bacterial overgrowth.

Treatment includes dietary modifications, medications to stimulate gastric emptying, medications to reduce vomiting, and surgical approaches. The prognosis in children, in idiopathic gastroparesis in adults, and in post-infectious is the likelihood of complete recovery; in diabetic gastroparesis it is more likely to be slowly progressive and even fatal. The prevalence of post-infectious gastroparesis is up to 4% of the population; up to 90% are young women.

<https://en.wikipedia.org/wiki/Gastroparesis> (accessed on 06.04.2020)

## **Text 9: Inflammation**

Inflammation (from Latin: inflammatio) is part of the complex biological response of body tissues to harmful stimuli, such as pathogens, damaged cells, or irritants, and is a protective response involving immune cells, blood vessels, and molecular mediators. The function of inflammation is to eliminate the initial cause of cell injury, clear out necrotic cells and tissues damaged from the original insult and the inflammatory process, and initiate tissue repair.

The five classical signs of inflammation are heat, pain, redness, swelling, and loss of function (Latin calor, dolor, rubor, tumor, and functio laesa). Inflammation is a generic response, and therefore it is considered as a mechanism of innate immunity, as compared to adaptive immunity, which is specific for each pathogen. Too little inflammation could lead to progressive tissue destruction by the harmful stimulus (e.g. bacteria) and compromise the

survival of the organism. In contrast, chronic inflammation is associated with various diseases, such as hay fever, periodontal disease, atherosclerosis, and osteoarthritis.

Inflammation can be classified as either acute or chronic. Acute inflammation is the initial response of the body to harmful stimuli and is achieved by the increased movement of plasma and leukocytes (especially granulocytes) from the blood into the injured tissues. A series of biochemical events propagates and matures the inflammatory response, involving the local vascular system, the immune system, and various cells within the injured tissue. Prolonged inflammation, known as chronic inflammation, leads to a progressive shift in the type of cells present at the site of inflammation, such as mononuclear cells, and is characterized by simultaneous destruction and healing of the tissue from the inflammatory process.

<https://en.wikipedia.org/wiki/Inflammation> (accessed on 06.04.2020)

## **Text 10: Infection**

An infection is the invasion of an organism's body tissues by disease-causing agents, their multiplication, and the reaction of host tissues to the infectious agents and the toxins they produce. An infectious disease, also known as a transmissible disease or communicable disease, is an illness resulting from an infection.

Infections are caused by infectious agents (pathogens) including:

Viruses and related agents such as viroids and prions

Bacteria

Fungi, further subclassified into:

Ascomycota, including yeasts such as *Candida*, filamentous fungi such as *Aspergillus*, *Pneumocystis* species, and dermatophytes, a group of organisms causing infection of skin and other superficial structures in humans.

Basidiomycota, including the human-pathogenic genus *Cryptococcus*.

Parasites, which are usually divided into:

Unicellular organisms (e.g. malaria, *Toxoplasma*, *Babesia*)

Macroparasites (worms or helminths) including nematodes such as parasitic roundworms and pinworms, tapeworms (cestodes), and flukes (trematodes, such as schistosomiasis)

Arthropods such as ticks, mites, fleas, and lice, can also cause human disease, which conceptually are similar to infections, but invasion of a human or animal body by these macroparasites is usually termed infestation. (Diseases caused by helminths, which are also macroparasites, are sometimes termed infestations as well, but are sometimes called infections.)

Hosts can fight infections using their immune system. Mammalian hosts react to infections with an innate response, often involving inflammation, followed by an adaptive response. Specific medications used to treat infections include antibiotics, antivirals, antifungals, antiprotozoals, and antihelminthics.

<https://en.wikipedia.org/wiki/Infection> (accessed on 06.04.2020)

## 9.2 Parsed Results in Python

### Text 1

Terms in Fred: 190

Terms in SDL: 61

Number of common terms: 21

List of common terms: *Adult-onset Diabetes, Autoimmune Response, Blood Sugar Level, Cardiovascular Disease, Diabetes mellitus, Diabetic Ketoacidosis, Frequent Urination, High Blood Sugar Level, Insulin, Insulin Resistance, Insulin-dependent Diabetes Mellitus, Juvenile Diabetes, Non Insulin-dependent Diabetes Mellitus, cardiovascular disease, cells, complications, diabetes, high blood sugar levels, hyperglycemic, hyperosmolar hyperglycemic state, metabolic disorders*

Jaccard index: 18.1

Evaluation of terms:

Precision = 21.88

Recall = 51.22

F1-score = 30.66

Relations in Fred: 529

Relations in SDL: 29

Number of common relations: 10

List of common relations:

{'Fred:': [['diabetes'], 'cause', ['complications']], 'SDL: ': [['Diabetes Mellitus', 'DM', 'Diabetes'], 'causalRelation ', ['complication', 'medical complication']]}

{'Fred:': [['cells'], 'respond', ['insulin']], 'SDL: ': [['cell'], 'activityRelation ', ['insulin']]}

{'Fred:': [['insulin-dependent diabetes mellitus', 'Insulin-dependent Diabetes Mellitus'], 'subClassOf', ['Diabetes Mellitus']], 'SDL: ': [['type 1', 'insulin-dependent diabetes mellitus', 'IDDM', 'juvenile diabetes'], 'genericRelation ', ['Diabetes Mellitus', 'DM', 'Diabetes']]}

{'Fred:': [['Adult-onset Diabetes', 'adult-onset diabetes'], 'subClassOf', ['Diabetes', 'diabetes']], 'SDL: ': [['type 1', 'insulin-dependent diabetes mellitus', 'IDDM', 'juvenile diabetes'], 'genericRelation ', ['Diabetes Mellitus', 'DM', 'Diabetes']]}

{'Fred:': [['Adult-onset Diabetes', 'adult-onset diabetes'], 'subClassOf', ['Diabetes', 'diabetes']], 'SDL: ': [['type 2 diabetes', 'T2D', 'adult-onset diabetes', 'non insulin-dependent diabetes mellitus', 'NIDDM'], 'genericRelation ', ['Diabetes Mellitus', 'DM', 'Diabetes']]}

{'Fred:': [['insulin-dependent diabetes mellitus', 'Insulin-dependent Diabetes Mellitus'], 'subClassOf', ['Diabetes Mellitus']], 'SDL: ': [['type 1', 'insulin-dependent diabetes mellitus', 'IDDM', 'juvenile diabetes'], 'genericRelation ', ['Diabetes Mellitus', 'DM', 'Diabetes']]}

{'Fred:': [['insulin-dependent diabetes mellitus', 'Insulin-dependent Diabetes Mellitus'], 'subClassOf', ['Diabetes Mellitus']], 'SDL: ': [['type 1', 'insulin-dependent diabetes mellitus', 'IDDM', 'juvenile diabetes'], 'genericRelation ', ['Diabetes Mellitus', 'DM', 'Diabetes']]}

{'Fred:': [['juvenile diabetes', 'Juvenile Diabetes'], 'subClassOf', ['Diabetes', 'diabetes']], 'SDL: ': [['type 1', 'insulin-dependent diabetes mellitus', 'IDDM', 'juvenile diabetes'], 'genericRelation ', ['Diabetes Mellitus', 'DM', 'Diabetes']]}

{'Fred:': [['cells'], 'respond', ['insulin']], 'SDL: ': [['cell'], 'activityRelation ', ['insulin']]}

{'Fred:': [['diabetes'], 'cause', ['complications']], 'SDL: ': [['Diabetes Mellitus', 'DM', 'Diabetes'], 'causalRelation ', ['complication', 'medical complication']]}

Jaccard index: 2.12

Accuracy = 34.48

Evaluation of relations:

Precision= 2.24

Recall = 28.57

F1-score = 4.16

## **Text 2**

Termini in Fred: 124

Termini in SDL: 57

Number of common terms: 11

List of common terms: *Double Vision, Immune System, Multiple Sclerosis, Multiple sclerosis, Muscle Weakness, Nervous System, Spinal Cord, Viral Infection, nervous system, signs, symptoms*

Jaccard index: 14.2

Evaluation of terms:

Precision= 19.20

Recall = 35.29

F1-score = 24.87

Relations in Fred: 261

Relations in SDL: 31

Number of common relations: 2

List of common relations:

```
{'Fred:': [['infection'], 'type', ['viral infection', 'Viral Infection']], 'SDL: ': [['infection'],  
'causalRelation ', ['viral infection', 'viral disease', 'infectious disease']]}
```

```
{'Fred:': [['infection'], 'type', ['viral infection', 'Viral Infection']], 'SDL: ': [['infection'],  
'causalRelation ', ['viral infection', 'viral disease', 'infectious disease']]}
```

Jaccard index: 0.8

Accuracy = 6.45

Evaluation of relations:

Precision= 0.90

Recall = 6.25

F1-score = 1.58

### **Text 3**

Termini in Fred: 156

Termini in SDL: 57

Number of common terms: 13

List of common terms: *Cancer, Colon Cancer, Colonoscopy, Inflammatory Bowel Disease, Invasive Surgery, Large Bowel, Ulcerative Colitis, anemia, diverticulosis, hematocrit, hemorrhage, malignancy, polyps*

Jaccard index: 15.23

Evaluation of terms:

Precision= 19.11

Recall = 42.86

F1-score = 26.43

Relations in Fred: 301

Relations in SDL: 36

Number of common relations: 2

List of common relations:

```
{'Fred:': [['colon cancer', 'Colon Cancer'], 'subClassOf', ['Cancer', 'cancer']], 'SDL: ':  
[['colorectal cancer', 'CRC', 'bowel cancer', 'colon cancer', 'rectal cancer'], 'genericRelation ',  
['cancer']]}
```

```
{'Fred:': [['colon cancer', 'Colon Cancer'], 'subClassOf', ['Cancer', 'cancer']], 'SDL: ':  
[['colorectal cancer', 'CRC', 'bowel cancer', 'colon cancer', 'rectal cancer'], 'genericRelation ',  
['cancer']]}
```

Jaccard index: 0.68

Accuracy = 5.56

Evaluation Relations:

Precision= 0.77

Recall = 5.41

F1-score = 1.35

#### **Text 4**

Termini in Fred: 142

Termini in SDL: 43

Number of common terms: 12

List of common terms: *Breech Birth, C-sections, Caesarean Delivery, Caesarean Section, Caesarean section, General Anaesthesia, Spinal Block, Umbilical Cord, Urinary Catheter, Vaginal Delivery, incisions, vaginal*

Jaccard index: 19.39

Evaluation of terms:

Precision= 22.54

Recall = 58.18

F1-score = 32.49

Relations in Fred: 326

Relations in SDL: 26

Number of common relations: 9

List of common terms:

{'Fred:': [['caesarean delivery', 'Caesarean Delivery'], 'useOf', ['surgery']], 'SDL: ': [['caesarean section', 'C-section', 'caesarean delivery'], 'genericRelation ', ['surgery', 'surgical procedure', 'operation']]}

{'Fred:': [['Vaginal Delivery', 'vaginal delivery'], 'hasQuality', ['Vaginal', 'vaginal']], 'SDL: ': [['vaginal delivery'], 'spatialRelation ', ['vagina']]}

{'Fred:': [['Urinary Catheter', 'urinary catheter'], 'drain', ['bladder']], 'SDL: ': [['urinary catheter'], 'spatialRelation ', ['bladder', 'urinary bladder']]}

{'Fred:': [['incision'], 'made', ['abdomen']], 'SDL: ': [['incision', 'pubic incision', 'Pfannenstiel incision'], 'spatialRelation ', ['abdomen']]}

{'Fred:': [['incision'], 'made', ['abdomen']], 'SDL: ': [['incision', 'pubic incision', 'Pfannenstiel incision'], 'spatialRelation ', ['abdomen']]}

{'Fred:': [['Urinary Catheter', 'urinary catheter'], 'drain', ['bladder']], 'SDL: ': [['urinary catheter'], 'spatialRelation ', ['bladder', 'urinary bladder']]}

{'Fred:': [['incision'], 'made', ['abdomen']], 'SDL: ': [['incision', 'pubic incision', 'Pfannenstiel incision'], 'spatialRelation ', ['abdomen']]}

{'Fred:': [['caesarean delivery', 'Caesarean Delivery'], 'useOf', ['surgery']], 'SDL: ': [['caesarean section', 'C-section', 'caesarean delivery'], 'genericRelation ', ['surgery', 'surgical procedure', 'operation']]}

{'Fred:': [['Vaginal Delivery', 'vaginal delivery'], 'hasQuality', ['Vaginal', 'vaginal']], 'SDL: ': [['vaginal delivery'], 'spatialRelation ', ['vagina']]}

Jaccard index: 2.86

Accuracy = 34.62

Evaluation Relations:

Precision= 3.07

Recall = 29.03

F1-score = 5.56

## **Text 5**

Termini in Fred: 126

Termini in SDL: 55

Number of common terms: 12

List of common terms: *Anti-inflammatory, Blood Clotting, Drug Class, Heart Attack, Kidney Disease, NSAID, Paracetamol, enzymes, mediators, prostaglandins, steroids, thromboxanes*

Jaccard index: 13.53

Evaluation of terms:

Precision= 18.25

Recall = 34.33

F1-score = 23.83

Relations in Fred: 266

Relations in SDL: 30

Number of common relations: 2

List of common relations:

{'Fred:': [['prostaglandins'], 'involved', ['inflammation']], 'SDL: ': [['prostaglandin', 'PG'], 'activityRelation ', ['inflammation']]}

{'Fred:': [['prostaglandins'], 'involved', ['inflammation']], 'SDL: ': [['prostaglandin', 'PG'], 'activityRelation ', ['inflammation']]}

Jaccard index: 0.78

Accuracy = 6.67

Evaluation Relations:

Precision = 0.87

Recall = 6.45

F1-score = 1.54

## **Text 6**

Termini in Fred: 167

Termini in SDL: 29

Number of common terms: 8

List of common terms: *Birth Control, Breakthrough Bleeding, Cervical Mucus, Deep Vein Thrombosis, Endocrine System, Male Hormonal Contraceptive, Selective Oestrogen Receptor Modulator, progestins*

Jaccard index: 10.27

Evaluation of terms:

Precision= 11.38

Recall = 51.35

F1-score = 18.63

Relations in Fred: 482

Relations in SDL: 21

Number of common relations: 0

List of common relations:

Jaccard index: 0.0

Accuracy = 0.0

Evaluation Relations:

Precision= 0.00

Recall = 0.00

F1-score = 0.00

### **Text 7**

Termini in Fred: 130

Termini in SDL: 57

Number of common terms: 14

List of common terms: *Depression, Dry Mouth, Emotional Lability, Major Depressive Disorder, Serotonin, Serotonin Syndrome, Weight Gain, addictions, antidepressants, drugs, norepinephrine, serotonin toxicity, sexual dysfunction, symptoms*

Jaccard index: 16.86

Evaluation of terms:

Precision= 22.31

Recall = 40.85

F1-score = 28.86

Relations in Fred: 284

Relations in SDL: 37

Number of common relations: 2

List of common relations:

{'Fred.': [['symptoms'], 'mostProminentSymptomOf', ['MDD']], 'SDL: ': [['symptom'], 'activityRelation ', ['major depressive disorder', 'MDD', 'depression']]}

{'Fred.': [['symptoms'], 'mostProminentSymptomOf', ['MDD']], 'SDL: ': [['symptom'], 'activityRelation ', ['major depressive disorder', 'MDD', 'depression']]}

Jaccard index: 0.72

Accuracy = 5.41

Evaluation Relations:

Precision= 0.83

Recall = 5.26

F1-score = 1.43

## **Text 8**

Termini in Fred: 182

Termini in SDL: 78

Number of common terms: 22

List of common terms: *Abdominal Bloating, Abdominal Pain, Autonomic Neuropathy, Barium Swallow, Complications, Diabetes Mellitus, Diabetic Gastroparesis, Diagnosis, Digestive Tract, Gastric Emptying, Gastric-emptying Scan, Gastroparesis, Idiopathic Gastroparesis,*

*Intestinal Obstruction, Nerve Damage, Stomach, Weight Loss, bacterial overgrowth, bezoars, gastric-emptying, medications, satiety*

Jaccard index: 22.61

Evaluation of terms:

Precision= 28.57

Recall = 52.00

F1-score = 36.88

Relations in Fred: 384

Relations in SDL: 44

Number of common relations: 6

List of common terms:

{'Fred:': [['medications'], 'reduce', ['vomiting']], 'SDL: ': [['medication', 'medicine', 'pharmaceutical drug', 'drug'], 'activityRelation ', ['vomiting', 'emesis']]}

{'Fred:': [['idiopathic gastroparesis', 'Idiopathic Gastroparesis'], 'subClassOf', ['gastroparesis']], 'SDL: ': [['idiopathic gastroparesis', 'gastroparesis with no known cause'], 'genericRelation ', ['gastroparesis', 'delayed gastric emptying']]}

{'Fred:': [['Diabetic Gastroparesis', 'diabetic gastroparesis'], 'subClassOf', ['gastroparesis']], 'SDL: ': [['diabetic gastroparesis'], 'genericRelation ', ['gastroparesis', 'delayed gastric emptying']]}

{'Fred:': [['Diabetic Gastroparesis', 'diabetic gastroparesis'], 'subClassOf', ['gastroparesis']], 'SDL: ': [['diabetic gastroparesis'], 'genericRelation ', ['gastroparesis', 'delayed gastric emptying']]}

{'Fred:': [['idiopathic gastroparesis', 'Idiopathic Gastroparesis'], 'subClassOf', ['gastroparesis']], 'SDL: ': [['idiopathic gastroparesis', 'gastroparesis with no known cause'], 'genericRelation ', ['gastroparesis', 'delayed gastric emptying']]}

{'Fred:': [['medications'], 'reduce', ['vomiting']], 'SDL: ': [['medication', 'medicine', 'pharmaceutical drug', 'drug'], 'activityRelation ', ['vomiting', 'emesis']]}

Jaccard index: 1.58

Accuracy = 13.64

Evaluation Relations:

Precision= 1.78

Recall = 12.77

F1-score = 3.12

## **Text 9**

Termini in Fred: 169

Termini in SDL: 60

Number of common terms: 11

List of common terms: *Adaptive Immunity, Biological Response, Cell Injury, Chronic Inflammation, Functio Laesa, Inflammation, Innate Immunity, Vascular System, inflammatio, pathogens, tissues*

Jaccard index: 18.23

Evaluation of terms:

Precision= 21.89

Recall = 52.11

F1-score = 30.83

Relations in Fred: 412

Relations in SDL: 35

Number of common relations: 2

List of common relations:

```
{'Fred:': [['inflammation'], 'type', ['Chronic Inflammation', 'chronic inflammation']], 'SDL: ':  
[['inflammation'], 'propertyRelation ', ['chronic inflammation']]}
```

```
{'Fred:': [['inflammation'], 'type', ['Chronic Inflammation', 'chronic inflammation']], 'SDL: ':  
[['inflammation'], 'propertyRelation ', ['chronic inflammation']]}
```

Jaccard index: 0.51

Accuracy = 5.71

Evaluation Relations:

Precision= 0.56

Recall = 5.56

F1-score = 1.01

## **Text 10**

Termini in Fred: 139

Termini in SDL: 70

Number of common terms: 26

List of common terms: *Babesia, Bacteria, Basidiomycota, Candida, Communicable Disease, Cryptococcus, Disease, Infection, Infectious Disease, Pneumocystis, Toxoplasma, flukes, helminths, hosts, immune\_system, infections, infestations, macroparasites, mites, nematodes, pathogens, roundworms, ticks, tissues, toxins, transmissible\_disease*

Jaccard index: 21.76

Evaluation of terms:

Precision= 30.22

Recall = 43.75

F1-score = 35.74

Relations in Fred: 351

Relations in SDL: 44

Number of common relations: 3

List of common relations:

```
{'Fred:': [['Infectious Disease', 'infectious disease'], 'subClassOf', ['Disease', 'disease']], 'SDL: ':  
[['infectious disease', 'transmissible disease', 'communicable disease'], 'genericRelation ',  
['illness', 'disease']]}
```

```
{'Fred:': [['Infectious Disease', 'infectious disease'], 'subClassOf', ['Disease', 'disease']], 'SDL: ':  
[['infectious disease', 'transmissible disease', 'communicable disease'], 'genericRelation ',  
['illness', 'disease']]}
```

```
{'Fred:': [['Communicable Disease', 'communicable disease'], 'subClassOf', ['Disease',  
'disease']], 'SDL: ': [['infectious disease', 'transmissible disease', 'communicable disease'],  
'genericRelation ', ['illness', 'disease']]}
```

Jaccard index: 0.87

Accuracy = 6.82

Evaluation Relations:

Precision= 0.99

Recall = 6.52

F1-score = 1.72

## Abstract (English)

This master's thesis investigated the potential of FRED, a Semantic Web machine-reading tool, for automatic term and relation extraction in terminology work applied in the biomedical domain. To achieve this objective, a corpus of ten English-language Wikipedia articles covering different medical topics was analysed. The concepts and relations extracted automatically by FRED were compared with concept system manually created in SDL Multiterm.

The quantitative analysis demonstrated that FRED is moderately effective in identifying relevant domain concepts. The system retrieved a substantial proportion of the concepts identified during the manual terminology analysis. The correspondence between relations extracted by FRED and those manually chosen in SDL are considerably lower. The qualitative analysis, on the contrary, reveals that this limited overlap is attributable to the different representational principles underlying the two approaches rather than to extraction errors.

FRED establishes semantic relations on the basis of linguistic and syntactic structure of the source text and formalises them through ontology-based representations. SDL terminology analysis, by contrast, seeks to organize domain knowledge and terminology by concept systems. As a result, the two approaches often identify similar concepts but represent their relationships in different ways.

By qualitative analysis several examples emerge, which demonstrate the potential of FRED. The tool is capable of identify semantic correspondences, recognise synonymous concepts and establish links to external knowledge resources such as DBpedia, even when the related concepts do not occur explicitly in the source text. On the contrary, the analysis also revealed several limitations, including the generation of duplicate relations and the extraction of relations that reflect linguistic structure rather than conceptual relation relevant for terminology work.

Overall, FRED can provide useful support during the initial stages of terminology work, particularly for concept identification and semantic enrichment. Nevertheless, automatic ontology extraction remains insufficient to replace manual terminology analysis. The results therefore support the use of a hybrid approach in which ontology extraction tools assist, rather than substitute, human terminologists.

## Abstract (German)

Diese Masterarbeit untersucht das Potenzial von FRED, einem Semantic-Web-Tool zur automatischen Textanalyse, für die automatische Extraktion von Konzepten und Relationen im Rahmen terminologischer Arbeit im biomedizinischen Bereich. Hierfür wurde ein Korpus aus zehn englischsprachigen Wikipedia-Artikeln zu unterschiedlichen Themen untersucht. Die von FRED automatisch extrahierten Konzepte und Relationen wurden mit den in SDL MultiTerm manuell erstellten Begriffssystemen verglichen.

Die quantitative Auswertung zeigt, dass FRED relevante Fachkonzepte in beachtlichem Umfang identifizieren kann. Ein Teil der Konzepte, die im Rahmen der manuellen Terminologearbeit erfasst wurden, konnte auch durch FRED erkannt werden. Deutlich geringer fiel dagegen die Übereinstimmung bei den Relationen aus. Die von FRED extrahierten Relationen werden aus der sprachlichen und semantischen Struktur eines Textes abgeleitet und werden in ontologiebasierten Strukturen formalisiert. Bei der Terminologearbeit steht hingegen die Modellierung von Begriffssystemen im Vordergrund. Beide Ansätze erfassen daher häufig ähnliche Konzepte, stellen die Beziehungen zwischen diesen Konzepten jedoch unterschiedlich dar.

Die qualitative Analyse zeigt, dass die Unterschiede nicht auf Extraktionsfehler zurückzuführen sind, sondern auf die unterschiedlichen Grundlagen der beiden Ansätze. Aber wurden mehrere Stärken von FRED sichtbar. Das System konnte semantische Zusammenhänge erkennen, synonymische Konzepte miteinander verknüpfen und Verbindungen zu externen Wissensressourcen, z. B. DBpedia, herstellen. Dies gelang teilweise sogar dann, wenn die entsprechenden Konzepte im Ausgangstext nicht ausdrücklich genannt wurden. Die Untersuchung zeigte jedoch auch einige Einschränkungen. So erzeugt FRED wiederholt doppelte Relationen und bildet häufig sprachliche Strukturen ab, die aus terminologischer Sicht nicht relevant sind.

Inbesondere in den frühen Phasen der Terminologearbeit kann FRED eine wertvolle Unterstützung darstellen, bei der Konzeptextraktion oder der semantischen Anreicherung der Terminologie. Für die Erstellung von Begriffssystemen reicht die automatische Extraktion von Ontologien jedoch nicht aus, um die manuelle Terminologearbeit zu ersetzen. Daher erscheint ein hybrider Ansatz sinnvoll, bei dem automatische Verfahren die Arbeit von Terminologinnen unterstützen, ohne deren fachliche Analyse zu ersetzen.

## **Declaration on the Use of AI-Enabled Tools**

In the preparation of the present thesis, AI-enabled tools were used exclusively for the linguistic and stylistic revision of the text.

The use of these tools was strictly limited to improving linguistic clarity, stylistic coherence, and the formal quality of the independently developed content.

Full responsibility for the content of this thesis lies with the author. The use of the aforementioned tools complied with the applicable guidelines of the Centre for Translation Studies (Zentrum für Translationswissenschaft) of the University of Vienna regarding good academic practice.