Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2021

Université Savoie Mont Blanc

3 & 4 juin 2021

Les ouvrages TOTh précédents sont disponibles :

- sur le site de l'Université Savoie Mont Blanc (btk.univ-smb.fr/livres)
- sur le site du Comptoir des Presses d'Universités (www.lcdpu.fr)
- ou auprès de : contact@toth.condillac.org

Éditeur : Presses Universitaires Savoie Mont Blanc
27 rue Marcoz
BP 1104
73011 CHAMBÉRY CEDEX
www.univ-smb.fr

Réalisation : C. Brun, C. Roche
Collection « Terminologica »
ISBN : 978-2-37741-079-8
ISSN : 2607-5008
Dépôt légal : juillet 2022

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2021

Université Savoie Mont Blanc

3 & 4 juin 2021

http://toth.condillac.org

avec le soutien de :

- Université Savoie Mont Blanc
- École d'ingénieurs Polytech Annecy Chambéry
- Ministère de la Culture. Ce projet est soutenu financièrement par le Ministère de la Culture - Délégation Générale à la Langue Française et aux Langues de France

MINISTÈRE
DE LA CULTURE
Liberté
Égalité
Fraternité

Presses Universitaires Savoie Mont Blanc
Collection « Terminologica »

Comité scientifique

Président du Comité scientifique : Christophe Roche

Comité de pilotage

Rute Costa	Universidade Nova de Lisboa
Humbley John	Université Paris 7
Kockaert Hendrik	University of Leuven
Christophe Roche	Université Savoie Mont Blanc

Comité de programme 2019

Le comité de programme est constitué chaque année à partir du comité scientifique de TOTh en fonction des soumissions reçues. La composition du comité scientifique est accessible à l'adresse suivante : http://toth.condillac.org/committees

Amparo Alcina	Universitat Jaume I – Spain
Xiaomi An	Renmin University – China
Albina Auksoriute	Institute of the Lithuanian Language – Lithuania
Jean-Paul Barthès	Université Technologie de Compiègne – France
Christopher Brewster	Maastricht University – Netherlands
Nicolleta Calzolari	Instituto di Linguistica Computazionale, CNR – Italy
Danielle Candel	CNRS, Université Paris Diderot – France
Sylviane Cardey	Université de Franche-Comté – France
Stéphane Chaudiron	Université de Lille 3 – France
Manuel Célio Conceição	Universidade do Algarve – Portugal
Rute Costa	Universidade NOVA de Lisboa – Portugal
Éric De La Clergery	INRIA – France
Luc Damas	Université Savoie Mont-Blanc – France
Dardo De Vecchi	Kedge Business School – France
Thierry Declerck	DFKI – Germany
Valérie Delavigne	Université Paris 3 – France
Sylvie Desprès	Université Paris 13 – France
Juan Carlos Diaz Vasquez	EAFIT University – Colombia
Pamela Faber	Universidad de Granada – Spain
Christiane Fellbaum	Princeton University – USA
Cécile Frérot	Université Stendhal Grenoble 3 – France
Iolanda Galanes	Universidade de Vigo – Spain
Teodora Ghiviriga	Alexandru Ioan Cuza University – Romania
Rufus Gouws	University of Stellenbosch – South Africa
Jean-Yves Gresser	ancien Directeur à la Banque de France – France
John Humbley	Université Paris 7 – France
Yangli Jia	University of Liaocheng – China
Kyo Kageura	University of Tokyo – Japan
Barbara Karsch	BIK Terminology – USA
Hendrik Kockaert	University of Leuven – Belgium

Hélène Ledouble	Université de Toulon – France
Patrick Leroyer	Aarhus University – Denmark
Georg Löckinger	University of Applied Sciences Upper Austria – Austria
John McCrae	National University of Ireland – Ireland
Candida Jaci de Sousa Melo	Universidade Federal do Rio Grande do Norte – Brazil
Christine Michaux	Université de Mons – Belgium
Fidelma Ní Ghallchobhair	Foras na Gaeilge, Irish-Language Body – Ireland
António Pareja Lora	Universidad Complutense de Madrid – Spain
Silvia Piccini	Italian National Research Council – Italy
Suzanne Pinson	Université Paris Dauphine - France
Maria Pozzi	El colegio de México – Mexico
Bihua Qiu	China National Committee for Terms in Sciences and Technologies – China
Jean Quirion	Université d'Ottawa – Canada
Renato Reinau	Université de Genève – Switzerland
Christophe Roche	Université Savoie Mont-Blanc – France
Mathieu Roche	CIRAD – France
Laurent Romary	INRIA & HUB-ISDL – Germany
Micaela Rossi	Università degli studi di Genova – Italy
Klaus-Dirk Schmitz	Cologne University – Germany
Frieda Steurs	University of Leuven – Belgium
Toma Tasovac	Belgrade Center for Digital Humanities – Serbia
Philippe Thoiron	Université Lyon 2 – France
Kara Warburton	City University of Hong Kong – China
Maria Teresa Zanola	Università Cattolica del Sacro Cuore – Italy

Avant-propos

La pandémie de COVID-19 a durablement impacté nos façons de travailler, et en particulier l'organisation d'événements tels que les conférences. Si rien ne peut remplacer la richesse des contacts humains que procure le présentiel, il nous faut accepter d'autres modes de participation et d'échanges. La participation à distance, l'enregistrement des interventions en font partie. Au-delà d'une gestion différente des coûts et du temps, cela offre d'autres perspectives dont une plus grande diffusion des travaux menés et donc une meilleure visibilité.

Nous avons pu revenir en 2021 à la traditionnelle planification de TOTh la 1re semaine de juin établie depuis 2007. La formation et la conférence se sont déroulées conjointement en présentiel et à distance, avec une très forte participation à distance, les restrictions sanitaires étant toujours en vigueur. L'organisation est certes plus compliquée et même si le présentiel devrait, nous l'espérons, revenir en force l'année prochaine, la participation à distance sera dorénavant proposée. L'Université de Savoie et l'équipe Condillac sur lesquelles reposent l'organisation de la conférence et la publication des actes seront plus fortement impliquées.

Avant de présenter les actes de cette année, j'aimerais remercier à nouveau les membres du Comité international de programme 2021 pour leur travail. Fortement mobilisés – les soumissions sont évaluées par au moins trois relecteurs – ils sont garants de la qualité des travaux menés à TOTh. Je rappelle que le Comité de programme est constitué chaque année à partir du Comité scientifique de TOTh en fonction des soumissions reçues. Le Comité scientifique est composé de 75 membres, experts internationalement reconnus du domaine, représentant 24 nationalités différentes.

La Conférence TOTh 2021 s'est ouverte avec la conférence invitée de notre collègue Nicola Guarino, bien connu des « ontologues », qui a dirigé le laboratoire d'ontologie appliquée de Trento rattaché au Conseil National de la Recherche italienne. Son intervention a porté sur « Events and their Names », un sujet aussi difficile qu'il est important. Nous présentons ici un résumé d'une communication que Nicola Guarino développera dans une communication ultérieure.

Notre collègue, François Gaudin, de l'Université de Rouen, a proposé cette année une Disputatio nous invitant à une lecture sociolinguistique de la référence chez Hilary Putnam.

Sur les 13 communications présentées, seules 10 ont été retenues pour publication. Elles ont abordé de nombreux sujets tant théoriques que pratiques portant sur des domaines aussi variés que les humanités numériques, la finance, les modèles de représentation, ou l'harmonisation de termes et de concepts.

Cette année nous avons eu le plaisir de décerner deux prix jeunes chercheurs. Cela est suffisamment exceptionnel pour que nous y consacrions quelques lignes. Instauré en 2011, ce prix n'a été décerné que deux fois, en 2011 et en 2018. Cette année ce sont deux jeunes chercheuses, toutes les deux italiennes, qui ont été récompensées. Cristina Farroni, de l'Università degli studi di Macerata, a présenté une contribution intitulée « Collaborative terminology management in a business environment : a case study in the field of wood paints and coatings ». Federica Vezzani, de l'Università degli studi di Padova, nous a présenté ses travaux en français sur le thème de « La gestion de (méta)données terminologiques « FAIR » : le répertoire de catégories de données de la ressource TriMED ».

Plus de 60 personnes ont suivi de manière assidue les présentations, ce qui correspond à la participation moyenne à la conférence. 21 pays étaient représentés : Afrique du Sud, Albanie, Allemagne, Autriche, Belgique, Chine, Espagne, États-Unis, France, Ghana, Grèce, Hongrie, Irlande, Italie, Lituanie, Luxembourg, Portugal, Roumanie, Royaume-Uni, Sénégal, Suisse.

Je vous invite à découvrir les communications que nous avons retenues à travers ces actes réalisés avec M^{me} Catherine Brun et publiés aux Presses Universitaires Savoie Mont-Blanc. Les actes des années précédentes sont accessibles à partir du site de la conférence (http://toth.condillac.org/) et des Presses Universitaires Savoie Mont Blanc (https://btk.univ-smb.fr/livres/?fwp_collections_revues=terminologica).

Avant de vous souhaiter bonne lecture, j'aimerais terminer en remerciant le Ministère de la Culture, et plus précisément la Délégation Générale à la Langue Française et aux Langues de France, l'Université Savoie Mont-Blanc, l'École Polytech Annecy-Chambéry et l'équipe Condillac pour leur support et leur aide financière à l'organisation de la conférence et à la publication des actes.

Christophe Roche
Président du Comité scientifique

Sommaire

Conférence d'ouverture 13

Events and their Names
Nicola Guarino 15

Articles 19

Formalizations of Knowledge and/in Semiotic Models in Terminology Science
Marija Ivanović, Thierry Declerck, 21

Creating a termino-ontological resource for translators in the domain of viral infectious diseases
Alice Sanfilippo 39

Crisis in troubled ancient times : ontological modelling of textual evidence from Greek historians
Maria Papadopoulou, Eleni-Melina Tamiolaki, Christophe Roche 57

Evolution of Modular Ontology : Application to Personalization
Rahma Dandan, Sylvie Despres 83

The Harmonization of Terms and Concepts in Function of the Standardization of Computer Terminology in the Albanian Language
Anila Çepani, Adelina Çerpja 105

Extending TBX2RDF
Thierry Declerck, Patricia Martín Chozas, Tom Winter, Tanja Wissik 123

Transformation of Graphical Semiotic Models into a Graph-based Formal Representation
Thierry Declerck, Marija Ivanović 137

Role of the Corpus in Ontoterminology : the Case of the Balance of Payments
Stéphane Carsenty 153

Prix jeunes chercheurs 179

Collaborative terminology management in a business environment: a case study in the field of wood paints and coatings
Cristina Farroni 181

La gestion de (méta)données terminologiques «FAIR»: le répertoire de catégories de données de la ressource TriMED
Federica Vezzani 199

Conférence d'ouverture

Events and their Names

Nicola Guarino

ISTC-CNR Laboratory for Applied Ontology, Trento, Italy
nicola.guarino@cnr.it

Abstract. In a well-known book, Bennett (1988) claimed that the semantics of event names ultimately depends on "local context and unprincipled intuitions". Here I will briefly report about some recent work that rejects this thesis, based on a novel ontological theory of events according to which the simplest events are qualitative changes, while most of the ordinary events described by natural language sentences are cognitively relevant clusters of co-occurring qualitative changes. According to this view, the lexicon provides systematic principles for individuating such clusters, classifying them into kinds, and imposing a specific structure on them, which reflects a cognitive focusing mechanism.

The title of this brief paper, based on the keynote I gave at the TOTh 2021 conference, is the same as that of a well-known book by Jonathan Bennett, where he claimed that the semantics of event names (mainly verbs and their nominalizations) ultimately depends on "local context and unprincipled intuitions". Here I will briefly report about some recent work that rejects this thesis, based on a novel ontological theory of events[1] whose main tenets are the following:

1. The simplest events are *qualitative changes*, i.e., changes *in a respect*. This means that there is a fundamental difference, in a change, between the object that changes (the participant in the change) and the actual subject of change.
2. Most of the events we refer to in our ordinary talk are cognitively relevant *clusters* of qualitative changes, on which we tend to impose a structure depending on how we perceive and we describe them.

[1] See Guarino, Baratella and Guizzardi (2022) for an extensive discussion.

3. Event kinds (typically lexicalized by verbs) provide criteria for isolating such clusters from the surrounding context and determining the structure imposed on them, distinguishing a *focus* which accounts for *what* happens, and an *internal context* which accounts for *how* it happens.

Claim 1 is based on a principle which was already present in Aristotle (Physics), and is at the root of Lombard's work (1986): events are changes in some respects. Our claim is that this principle should be taken seriously, by recognizing the ontological status of such "respects": they are qualitative aspects of objects, such as color or shape, which correspond to what DOLCE (Borgo & Masolo, 2009) and UFO (Guizzardi, 2005) call *individual qualities*, or just qualities for short. According to such notion of quality, the simplest events are not just changes in a quality, but changes of a quality in an object, in which the quality (directly or indirectly) inheres. Note that, differently from Lombard, we also consider a *stasis* as an event, assuming that in this case the quality value changes within a certain minimal threshold.

Claims 2 and 3 reflect the intuition that ordinary events are *cognitively constructed*: while single qualitative changes are of course independent from cognition, we tend to organize them in clusters that have a specific internal structure, accounting for a perception mechanism based on a *figure/ground* scheme (Talmy, 2000, p. 311).

Summing up, this work is based on two separate theories, a metaphysical one and a semantic one. Especially in the case of events, the metaphysical and semantic aspects are often highly intertwined. This is certainly true in Kim's account (1976), which has been indeed criticized by Bennett (1988), who accepted Kim's metaphysics but rejected his semantics. We believe however that also Kim's metaphysics should be revised, since defining events as property exemplifications unavoidably connects the nature of events to the way they are described. Abandoning such definition helps us to keep the two theories clearly separate, by recognizing first the nature of simple events as qualitative changes, and only then showing how language isolates specific clusters of simple events, and refers to their internal structure. As a result, we have a new fine-grained metaphysics of events that lies between the multiplicative and the unitarian approaches, and a semantic theory that, based on such metaphysics, provides a systematic account of the referential mechanisms of event nominals and event modifiers.

In our full paper we first discuss some evidence concerning how language refers to complex events, allowing them to be incrementally described by different kinds of modifiers. In particular, we discuss some challenges to Davidsonian compositionality of modifiers, which motivate a distinction between external and internal modifiers, along the lines suggested by Maienborn (2003). Clarifying the semantics of such distinction raises important questions concerning the internal structure of events and their relationship with the surrounding context, which are discussed in the rest of the paper.

References

Aristotle. Physics, book vi. In R. McKeon (Ed.), *The Basic Works of Aristotle.* Random House.

Bennett, J. (1988). *Events and their names*. Oxford University Press.

Borgo, S., & Masolo, C. (2009). Foundational choices in DOLCE. In *Handbook on ontologies* (pp. 361-381). Springer.

Guarino, N., Baratella, R., & Guizzardi, G. (2022). Events, their names, and their synchronic structure. *Applied Ontology, 17, 249-283.*

Guizzardi, G. (2005). *Ontological foundations for structural conceptual models*. CTIT; Centre for Telematics and Information Technology, University of Twente

Kim, J. (1976). Events as property exemplifications. In *Action theory* (pp. 159-177). Springer.

Lombard, L. B. (1986). Events: A metaphysical study.

Maienborn, C. (2003). Event-internal modifiers: Semantic underspecification and conceptual interpretation. In E. Lang, C. Maienborn, & C. Fabricius-Hansen (Eds.), *Modifying adjuncts.* Mouton de Gruyter.

Talmy, L. (2000). *Toward a cognitive semantics, volume 1: Concept structuring systems*. MIT Press.

ARTICLES

Formalizations of Knowledge and/in Semiotic Models in Terminology Science

Marija Ivanović*, Thierry Declerck**/***,

*University of Vienna
Centre for Translation Studies
Gymnasiumstraße 50
A-1190 Vienna, Austria
marija.ivanovic@univie.ac.at
**DFKI GmbH
Multilinguality and Language Technology Lab
Stuhlsatzenhausweg, 3
D-66123 Saarbrücken, Germany
declerck@dfki.de
https://www.dfki.de/~declerck /
***Austrian Academy of Sciences
Austrian Centre for Digital Humanities and Cultural Heritage
Sonnenfelsgasse, 19
A-1010 Vienna, Austria
declerck@dfki.de

Abstract. The aim of this study is to compare the semiotic models by Roche (2007) and Felber (1993): both deal with the formalization of knowledge, but were based on different theoretical influences, and traditions of terminology as well as in different times. While Felber's approach never was operationalised, Roche's semiotic model is the basis for the multilingual ontoterminology editor Tedi.

Both approaches are introduced, and then their perspective on natural language, concepts and concept relations, the role of logic, and formalized representation are compared. To encompass all these aspects and connect them to a real-world application, Tedi is used as a structure to which Felber's ideas are mapped. The analysis shows that Roche's and Felber's approach differ in their perspective on natural language, but deal both with concepts and concept relations, and use logic for the inheritance of characteristics in hierarchical concept structures. This could be a starting point for further analysis of those approaches and a possible combination.

1. Introduction

The semiotic triangle is one of the basic models in terminology science. It analyses the connections between the elements object, concept and sign. The semiotic triangle is generally expandable in terminology science (Wüster 1959), and in linguistics (Heger 1964; Melnikow 1988) as Wang (2016) shows. It is also applicable to different scenarios and developments in terminology science and ontology, as the work of Roche (2007), Felber (1993) and Sowa (2000) shows.

The perspectives of Roche (2007) and Felber (1993) on the semiotic triangle both include aspects of (logical) formalization and terminology, as well as representation, but in different degrees and forms as these two authors have different backgrounds: Roche in AI, and influences by de Saussure, and Felber by the Vienna school of terminology, and influences by Carnap and Wittgenstein. This paper wants to contribute to the tradition of comparing different schools of terminology, e.g. in Budin *et al*. (2006) and Laurén and Picht (1993), by analysing the common and distinguishing characteristics of the semiotic models of by Roche and Felber.

Roche (2007) offers two models (see Figure 1 and Figure 2), which consist of two semiotic triangles each. Basis for these models is the differentiation between language in use (langue d'usage), epistemological aspects of language (langue d'intellection) and language of representation (langage de représentation). Roche argues that because of this distinction the nature of the constituent parts of the two semiotic triangles in his models is different. The triangle which evolves from language in use (highlighted in blue in Figure 1 and Figure 2) consists of the elements: signifier, signified and praxis, and can be applied to linguistics in general in the first model and more specifically to language for special purposes (LSP) in the second model. For the second triangle in both models (highlighted in red in Figure 1 and Figure 2) strict epistemological aspects and principles are applied, and a formalized language of representation is used.

Felber (1993) takes a different approach: he uses the first of Wüster's (1959) semiotic quadrangles as a basis for a semiotic model of propositions (see Figure 3). In his model, several objects and their relations are abstracted into statements or formulae[1] in predicate logic (logischer Satz), which are built from concepts, and are then connected to a sentence built from signs for con-

[1] Here the term *statement* instead of *formula* should be used for better readability.

cepts, which manifests at an object level as a proposition (Aussage) in natural language and has to be standardized to be unambiguous.

Although both approaches include aspects of terminology and formalization, as well as natural or general language, they are structured differently. Furthermore, Roche's model is the basis for the ontoTerminology EDItor Tedi[2], while Felber's model never was operationalized. The aim of this paper is therefore to analyse the intersections of the models by Roche and Felber, and see, if and how Felber's model of the logical sentence could be mapped to Tedi. One of the difficulties in this attempt is the fact that the relevant articles were published in different times and languages, are influenced by different linguistic and philosophical traditions, and therefore use different terminologies.

2. Terminology and logical reasoning – two approaches

As a first step we will describe the models by Roche and Felber in all due detail to be able to compare them.

2.1. Roche: different aspects of language as the basis for two models

According to Roche, language has several relevant aspects and plays several roles when looking at the knowledge of a scientific or technological community. Language in use (langue d'usage), even when it is the language of a specialized community, reveals itself through the scientific or technical discourse of this community, which is based on texts. As the praxis of each field of discourse is central, the speaker, his or her intention as well as the possibility of interpretation play an important role. The extraction of concepts and concept systems from texts is possible, but these concept systems are usually not completely defined, structured or modelled (Roche 2007, 5) because they are a result of this fluid praxis of discourse. This aspect of language makes in Roche's first model (Figure 1) the linguistic semiotic triangle on the left necessary, which connects the signifier, the signified and the praxis of discourse. The signifier seems to evolve from the praxis of discourse and the complex communicative relations inherent to it. It therefore lacks the stability of the concept, which is the result of clearly defined epistemological principles (Roche 2007, 7). These epistemological principles are the basis for the definition of concepts and the structuring of concept systems, as well as the model-

[2] Detailed Information on Tedi can be found on http://ontoterminology.com/tedi.

ling of the objects of the world (Roche 2007, 5ff.) and can be based on different theoretical foundations. In terminology science this epistemological aspect of language is based on the analysis of the concept and its characteristics.

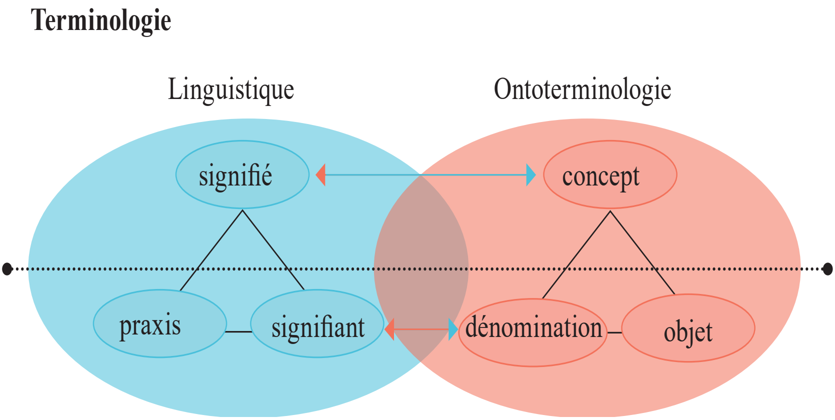


Fig. 1 – *Roche's two semiotic triangles for the realm of terminology (modified by us, and based on Roche 2007, 7)*

When these epistemological principles are combined with a language of representation (langage de représentation), they form the basis for the right triangle, which is used as a semiotic model for what Roche calls ontoterminology.

Ontoterminology is defined as

> «Une approche où l'ontologie joue un rôle fondamental à double titre: pour la construction du système notionnel et pour l'opérationnalisation de la terminologie. L'*ontoterminologie* insiste d'une part sur l'importance des principes épistémologiques qui président à la conceptualisation du domaine – c'est l'ontologie dans sa définition première –, et d'autre part sur la nécessité d'une approche scientifique de la terminologie où l'ingénieur joue un rôle fondamental – c'est l'ontologie dans ses définitions plus récentes». (Roche 2007, 8)

The language of representation is used to represent the concepts and the concepts system in a formalized way: e.g., by using ontology languages. A clear definition of concepts and concept systems on the one hand, and their representation based on axioms and rules, makes it possible to reduce the ambiguity inherent in the language in use. The representation based on a formal ontology language also makes concepts and concept system shareable and machine-readable (Roche 2007,6)

Roche (Figure 2) modified this first model, developed it further and in his second model describes the realm of ontoterminology, as a specific way of looking at concepts from two perspectives - one referring to the linguistic aspect in LSP and the other concentrating on ontologies, which are both an element of ontoterminology. The ontological aspect is analysed in the left triangle of this model, which consists of the concept, the object and the identifier. This triangle includes clearly defined epistemological principles as the basis for building concepts, and encompasses aspects of formalized and machine-readable structuring and representation of concepts. The triangle on the right includes aspects of language in use as it manifests in LSP within a certain community and through their praxis of discourse. The right triangle here is therefore a specific application of the linguistic triangle in Roche's first model (Figure 1), where the speaker, the intention behind the utterance, the unsaid, as well as the possibility of interpretation are central.

The traditional application of terminology as well as typical linguistic aspects are included in the triangle on LSP, while the degree of formalization and reusability is higher in the triangle referring to ontologies. What is specific for Roche's second model is the combination of strictly formalized ontological aspects as well as linguistic and social aspects as they arise in LSP to form an approach to concepts and concept systems which includes elements necessary in different settings of communication: between only humans, humans and machines and between machines. With this second model Roche shows the possibilities ontologies as well as LSP offer for terminology work.

Ontoterminologie

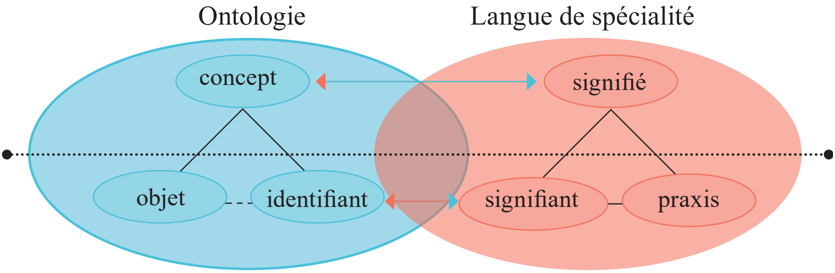


FIG. 2 – *Roche's model for ontoterminology (modified by us, and based on Roche 2007,13)*

2.2. Felber: a semiotic quadrangle for building logical statements

Felber's (1993, 98) model (see Figure 3) is based on Wüster's first semiotic quadrangle (1959). Wüster analyses the connection between objects, concepts, sign concepts and their manifestation in reality. For Felber concepts represent segments of reality, they are elements of thought (Denkgebilde).

Felber uses the structure of Wüster's quadrangle and applies it to analyse the relations between several objects in reality (Sachverhalte – bottom right area) – which are then abstracted into logical statements (logischer Satz – top right area) – a sequence of sign concepts representing them (Satz aus Begriffszeichen – top left area) and their manifestation as propositions in language (bottom left). These propositions are built from signs connected by a natural language syntax (Felber 1993, 81). The upper half of the model belongs to the realm of concepts (Begriffsebene) and the lower to the realm of objects (Gegenstandsebene).

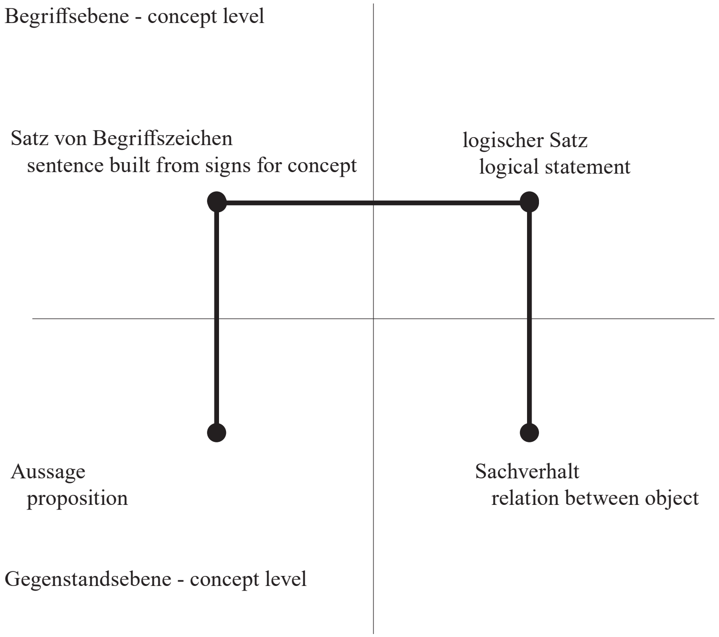


FIG. 3 – *Felber's semiotic quadrangle (modified by us, and based on Felber1993,98)*

This semiotic model is one of the elements, which form the theoretical basis for Felber's Wissenstechnik, a form of knowledge technology, which supports a possible form of logical reasoning, and is based on classical logic (including predicate logic). Felber's knowledge technology has its starting point in the logical statement, which is a unit of knowledge (Wissenseinheit). The logical statement uses predicate logic as a means for the representation, of relations between objects in reality (Sachverhalt) at the conceptual level. <<Gold is a metal>> is a relation between objects. The objects and their relations can be abstracted to a concept level in the form of a logical statement (Felber, 2001, 108). A logical statement can be true or false. Of the ontological reality (bottom right) it is referring to, it can only be said that it exists, while it can be referring to concrete or abstract occurrences (Felber 1993,68). Logical statements are used for logical reasoning, be it done by a human or as a machine, as Felber states (Felber 1993, 69). The connection between logical statement and proposition must be "adequate" (Felber 1993,99). As the proposition can manifest in different natural languages, different syntactic and grammatical means can be applied. According to Felber this leads to confusion because languages, which are not standardized (Gemeinsprachen) use signs and syntax in such a way that several interpretations of the same signs within a sentence are possible. To avoid this confusion, terminological standardization and standardization of syntax are necessary (Felber 1993, 98f.).

3. Comparison of the models

There are several aspects the models by Roche and Felber refer to: natural language, the world of concepts and concept relations, logic and formalized or standardized language: their approaches to these aspects will now be compared. To analyse the intersections of these models, the few examples of logical sentences and syllogisms Felber provides will be mapped to the structure of Tedi, the multilingual ontoterminology editor, which is based on Roche's distinction between the complementary linguistic and ontological dimension of ontoterminology (Figure 2). Tedi uses this distinction to build multilingual ontoterminologies, where the conceptual side of the terminology is structured and defined as a formal ontology, to which the terms in different languages are linked. Tedi therefore has a concept editor (ontological side) and a term editor (linguistic side). The former defines concepts and structures concept systems in a formal ontology. It is connected to the term editor, where the terms and their natural language definitions in different languages can be found. Felber's theoretical considerations on the other hand, so far have not been applied, and Felber (2001) offers only in a later publication fragments of the system he is envisioning. Nevertheless, these fragments can be used as a starting point for comparing his approach and the manifestation of Roche's in Tedi.

There are four main points that must be addressed:

- What is the perspective on natural languages of the two authors?
- Where can Felber's logical sentence as a form of concept relations be found in Tedi?
- Which role does logic play in both approaches?
- Which form of formalized representation is used in both approaches?

3.1. Natural languages

Roche and Felber both see a difference between general or natural language and standardized language or language of representation (in Roche's case). But they approach it in different ways: Roche argues that natural language in its everyday form, as well as LSP for a scientific or technical community, has a richness in its possibilities of expression, is complementary to standardized languages of representation and the strict epistemological principles they are based on. Natural language has certain characteristics (the importance of the speaker, the intention behind something said, the unsaid) and with this offers a richness of expression and possibilities. This aspect is visible in both of Roche's models and is a possible resource for formalized languages: new concepts develop through communication within a community of discourse and can then be formalized. This community is also able to verify the structure of formalized language and develop it further in this way.

Felber, on the other hand, sees the necessity to standardize natural languages to avoid ambiguity, and does not see it as much as a resource as Roche does. His model concentrates on the standardization of language, which is supposed to make it unambiguous.

3.2. Concepts and concept relations

When comparing Roche's triangle which is concerned with formalization (Figure 1 triangle on ontoterminology, and Figure 2 triangle on ontology), they have a lot in common with Felber's quadrangle, as Figure 4 and Figure 5 show. All three semiotic models refer to a connection between reality (represented by the object in Roche's models and relations between objects in Felber's model), its abstraction into concepts (Roche) and logical statements (Felber) and their representation in a formalized way by a designation/identifier in Roche's triangles and a proposition in a standardized language in Felber's case. The difference between the structure of the models by Roche and by Felber is that Felber includes a fourth element in his quadrangle: the sentence built from sign concepts in the top left area. If this element was excluded, both models would have the same structure. Felber (2011, 115) himself, in a later publication, reduces his semiotic quadrangle to this triangle "for the sake of simplicity".

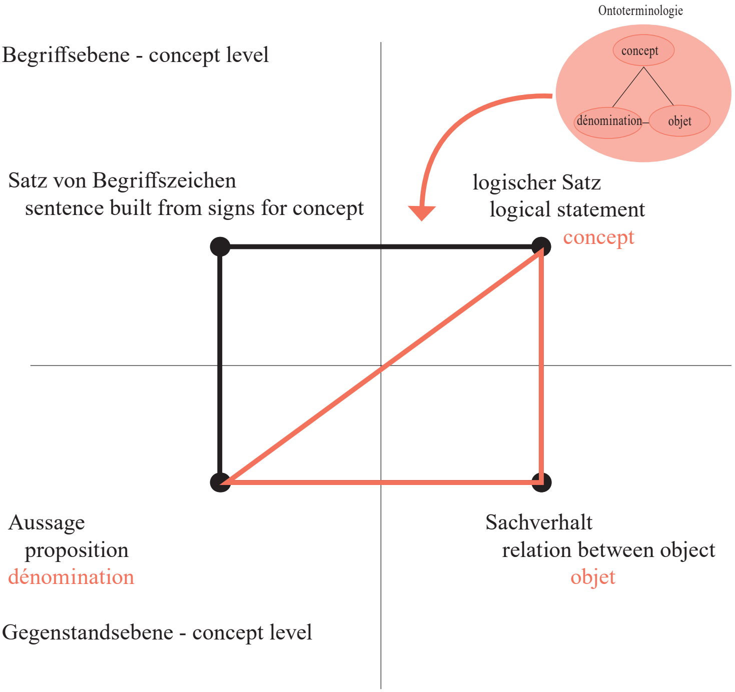


Fig. 4 – *Felber's semiotic quadrangle (1993,98) and Roche's triangle on ontoterminologie combined*

It is obvious that Roche uses his triangle to look at single concepts, but the relation between concepts is a basic aspect of terminology work. Therefore, concept relations can also be found in Tedi. The concept editor in Tedi does not only analyse concepts and their essential and differentiating characteristics. Tedi also structures concept systems using different concept relations. The main relations are generic (is-a) and partitive relations, but it is also possible to use and define other relations. Here a closer look will be taken at the generic relation because there is a parallel to one of Felber's approaches: Concepts and their relations for Felber are the building blocks of his logical sentences. In his publication from 2001 Felber provides an example for a relation between objects (indicated by double brackets) <<Metall ist ein Stoff>> (engl. <<metal is a substance>>), which is abstracted to a logical sentence, describing the (generic) relation between two concepts (indicated by single brackets) <Metall

ist ein Stoff> (engl. <metal is a substance>). In the example ontoterminology for seats by Roche (Figure 6) this form of generic relation can be found on the left side of the concepts editor – where the concept <seat> and its possible subordinate concepts can be found. In this hierarchy the concepts are designated using a concept name (the identifier in Figure 2), which includes the generic concept, and the inherited as well as differentiating characteristics.

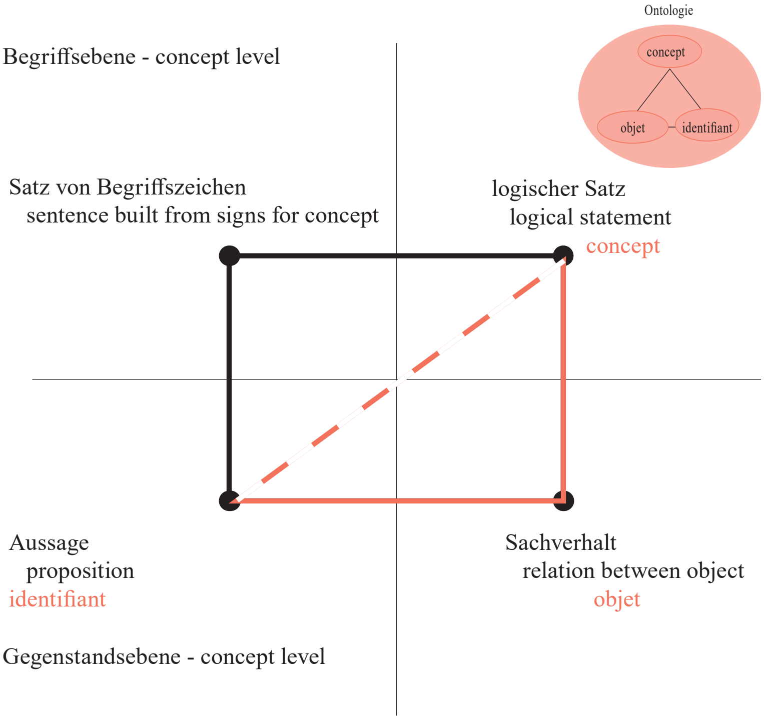


Fig. 5 – *Felber's semiotic quadrangle (1993,98) and Roche's triangle for ontology combined*

The specific concepts of the hierarchy are connected to terms from the term editor (Figure 6): The concept <Seat with feet for one person without arms without back> has the formal definition: <Seat with feet for one person> + /without arms/ + /without back/. <Seat with feet for one person> is here the generic concept to which the differentiating characteristics are added. The resulting concept is connected via the term editor to the English term *stool*, with its definition in natural language: Seat for one person, with feet, without arm and back. Translated to Felber's approach the hierarchical relation found here would have been formulated as the logical sentence: <A stool is a seat>. Felber's logical sentence can be found in Tedi's concept editor in the relations between the concepts. Felber did not say much about the relations he was envisioning. The example of the generic relation is just one possible relation. Tedi offers also other relations, such as the partitive relation as well as several others ('relatedTo'; 'hasFunction'; 'equivalentTo'; 'madeOf'; 'sequential'; 'causal; 'dependentOf'), and has the possibility to define new relations (Roche 2019, 26).

The formal definition of concepts is necessary for Roche's as well as Felber's approach, because Felber's knowledge technology was based on concepts and their constituting parts (Felber 2001,5). That Roche and Felber share some ideas on the formalization of concepts and concept systems, here again becomes visible (Figure 6, p. 34-35).

3.3. The role of logic

Predicate logic is the basis for Felber's logical sentence and was used widely in computational linguistics when he developed his model. Predicate logic, or at least a subset of it – description logic – was also the basis for some of the first modelling techniques of ontologies (Gómez-Pérez, Fernández-López, and Corcho 2004, 9). Today, ontology languages have in their detailed theoretical basis and elaboration, and also in their application been developed further from Felber's early theoretical considerations. Roche considers ontology languages a possible tool to model terminological concepts and concept structures. Therefore, there exists a connection between Roche's and Felber's approach. It is worth looking at how both authors utilized logic in their respective approaches.

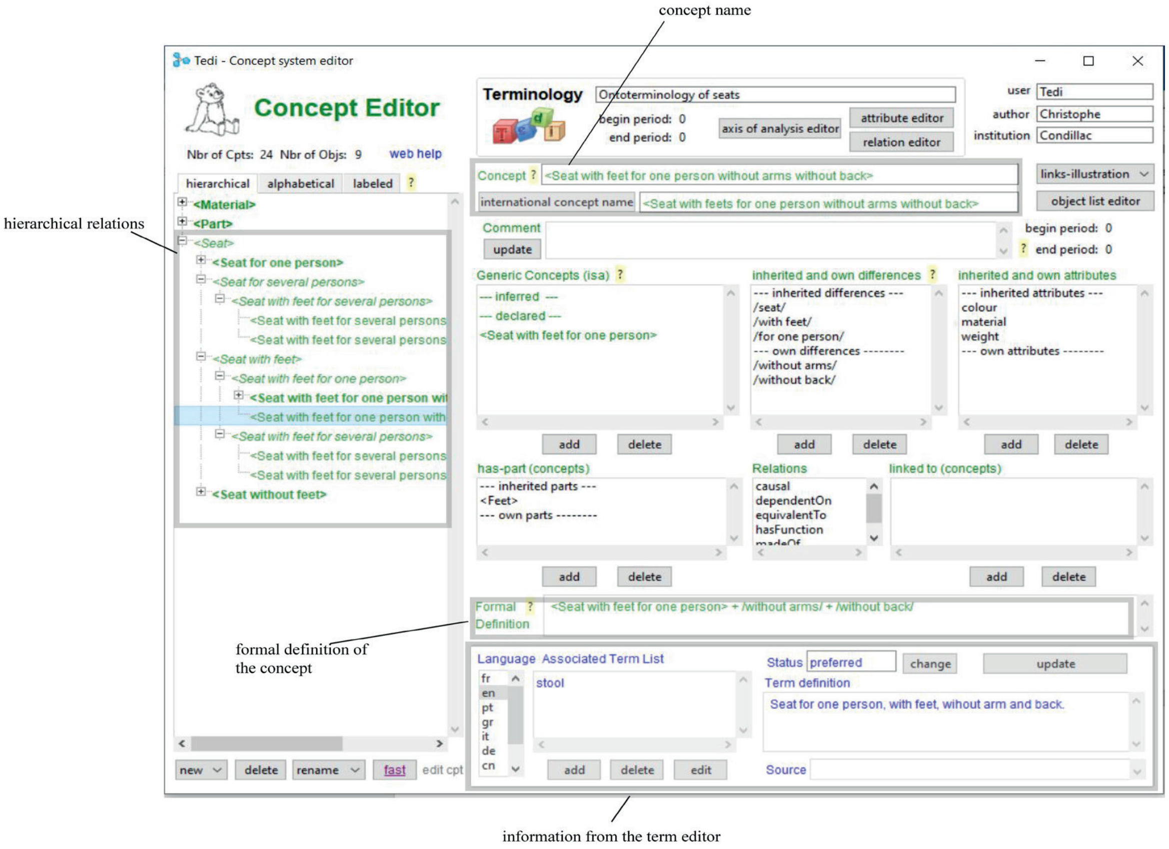
When it comes to logical aspects, Tedi uses axes of analysis to control the structure of the (generic) concept system and the inheritance of characteristics within it. In the seat-ontoterminology some of the axes of analysis would be *with/without feet*; *for one person/for several persons*; *with/ without back; with/ without arms*. These axes of analysis in Tedi are used to determine which characteristics are essential for structuring the concept system. This enables on the one hand to control the inheritance of characteristics when a subordinate concept is added, and on the other hand it is used to analyse where certain concepts, which already have a subordinate concept, can be placed within the concept system. The concept <seat with feet for one person without arms without back> is a subordinate concept to <seat with feet for one person>. It inherits the characteristics /with feet/ and /for one person/ and has own characteristics: /without arms/ and /without back/. The differentiating characteristics within one axis of analysis are exclusive of each other. Tedi controls this, and suggests only the possible characteristics, when characteristics are added.

Here it becomes obvious that concepts in this ontoterminological approach are part of concept relations, and therefore abstractions od relations between object in reality. This offers a parallel with one of Felber's fragments (2001,108), in which he provides the example how several relations between objects abstracted to logical sentences (consisting of concepts and their relations) could be connected into a chain of logical sentences, and would become a syllogism in logic:

<Gold ist ein Metall>	<gold is a metal >
<Metall ist ein Stoff>	<metal is a chemical substance >
<Gold ist ein Stoff>	<gold is a chemical substance>

This syllogism is built from concepts which are related to each other by a is-a or generic relation and can be translated into a hierarchical concepts structure that can be found in Tedi in the left vertical field of the concept editor. The chemical substance would be the first generic concept, with the other ones as subordinate concepts: <chemical substance> – <metal> – <gold>. In this structure <gold> in Tedi would inherit the characteristics of <metal>, as well as the characteristics of <chemical substance>. This process of inheriting characteristics is necessary for the form of reasoning Felber envisioned. Concepts here are building blocks of propositions and syllogisms. Therefore, the term logic found in Tedi can be considered a part of Felber's vision of the logical sentence, which uses classical logic, including term logic, predicate and propositional logic (Felber 2001, 17). On the other hand, the different relations Tedi offers can be considered as forms of predicate logic, as well as the basis for propositional logic. The use of logic is therefore another aspect that connects the approach by Roche and Felber's vision.

As some intersections of Roche's and Felber's thoughts can also be found in Tedi, Tedi could be considered the realisation of some of Felber's ideas, although time, background and tradition separate these authors.


concept name
hierarchical relations
formal definition of the concept
information from the term editor

Tedi - Concept system editor

Concept Editor

Nbr of Cpts: 24 Nbr of Objs: 9 web help

hierarchical | alphabetical | labeled | ?

- <Material>
- <Part>
- <Seat>
  - <Seat for one person>
  - <Seat for several persons>
    - <Seat with feet for several persons>
      - <Seat with feet for several persons
      - <Seat with feet for several persons
  - <Seat with feet>
    - <Seat with feet for one person>
      - <Seat with feet for one person wi
      - <Seat with feet for one person with
    - <Seat with feet for several persons>
      - <Seat with feet for several persons
      - <Seat with feet for several persons
  - <Seat without feet>

new | delete | rename | fast | edit cpt

Terminology: Ontoterminology of seats

begin period: 0
end period: 0

axis of analysis editor | attribute editor | relation editor

user: Tedi
author: Christophe
institution: Condillac

Concept ? <Seat with feet for one person without arms without back>

international concept name: <Seat with feets for one person without arms without back>

links-illustration | object list editor

Comment | update

begin period: 0
? end period: 0

Generic Concepts (isa) ?
--- inferred ---
--- declared ---
<Seat with feet for one person>

add | delete

inherited and own differences ?
--- inherited differences ---
/seat/
/with feet/
/for one person/
--- own differences --------
/without arms/
/without back/

add | delete

inherited and own attributes
--- inherited attributes ---
colour
material
weight
--- own attributes --------

add | delete

has-part (concepts)
--- inherited parts ---
<Feet>
--- own parts --------

add | delete

Relations
causal
dependentOn
equivalentTo
hasFunction
madeOf

linked to (concepts)

add | delete

Formal ? Definition: <Seat with feet for one person> + /without arms/ + /without back/

Language: fr, en, pt, gr, it, de, cn

Associated Term List: stool

add | delete | edit

Status: preferred | change | update

Term definition: Seat for one person, with feet, wihout arm and back.

Source

FIG. 6 – *A screenshot from the Seats-ontoterminology by Roche in Ted*

3.4. Representation in formalized languages

Tedi is able to export and therefore represent the concept systems in CVS, RDF/OWL, HTML and JSON in a formalized way (Roche 2019, 68ff.). Felber on the other hand, never operationalized his ideas and only left comments of what he was envisioning: formalizing LSP by using strict syntactical rules as well as unambiguous terminology and signs within the sentences. This was inspired by the clear syntactical rules and unambiguous use of signs and terminology he saw in mathematical logic and programming languages. (Felber 2001,111).

4. Conclusion

The two models by Roche differ in one important aspect from Felber's model: Roche chooses to include in his perspective on terminology (Figure 1) and ontoterminology (Figure 2) the linguistic dimension, with the intention of the speaker, the unsaid and the praxis of discourse in addition to the formalized aspect of language, which is the result of the application of strict epistemological principles as well as of rules for representation. Felber only analyses the formalized and rule-governed aspect of language in his model for knowledge technology.

A look at the epistemological aspects of the models by both authors showed that they have a lot in common: Both analyse objects or a combination of objects and their relations on the object level, which are then abstracted to a cognitive level as concepts or logical statements built from concepts, and are then represented in a formalized way. The only differences seem to be the structure of what they are analysing, whether they are concerned with single objects and concepts, or several connected objects which are abstracted to logical statements. But this does not mean that their approach is fundamentally different, on the contrary: the analysis of concepts and their relations is a basic element in terminology science and work, and also in Tedi. The hierarchical example relation Felber offers, can (in addition to several other relations) be found in Tedi. Therefore, the analysis of concepts is one building block of Tedi, just as it is a building block of Felber's knowledge technology.

The use of logic is another element that connects the approaches by Roche and Felber: The operationalized model in Tedi uses term logic to ensure the proper inheritance of characteristics between concepts and the correct structure of the concept system. This process seems to be a characteristic of syllogisms in logic, which Felber's approach and Tedi have in common. On the other hand, the concept relations in Tedi can be considered propositions that could be used in classical logic.

When it comes to the representation of the concepts or logical statements Felber's model stays theoretical, wishing for the standardization of language in a way programming languages are standardized, which might be just what Tedi is leaning towards, as it offers several possibilities of export and more or less formalized representation, such as CVS, RDF/OWL serializations and JSON.

This study is a first attempt to compare the models by Roche and Felber, authors from different times and backgrounds. Future research has to deepen the comparison between Roche's and Felber's approach and analyse other concept relations using the semiotic triangle. Furthermore, the theoretical background of both authors could be analysed to find the different and common theoretical influences.

References

Budin, Gerhard, Christer Laurén, Heribert Picht, Nina Pilke, Margaret Rogers, and Bertha Toft, eds. 2006. *The Theoretical Foundations of Terminology Comparison Between Eastern Europe and Western Countries. Proceedings of the Colloquium Held on 18 August 2003 in Surrey, Guilford, UK, in Conjunction with the 14th European Symposium on Language for Special Purposes.* Content and Communication. Terminology, Language Resources and Semantic Interoperability. Würzburg: Ergon. https://ubdata.univie.ac.at/AC05323397.

Felber, Helmut. 1993. *Allgemeine Terminologielehre und Wissenstechnik: Theoretische Grundlagen*. Wien: TermNet.

Felber, Helmut. 2001. *Allgemeine Terminologielehre, Wissenslehre und Wissenstechnik. Theoretische Grundlagen und philosophische Betrachtungen.* 3rd ed. Wien: TermNet - Internat. Network for Terminology.

Heger, K. 1964. 'Die Methodologischen Voraussetzungen von Onomasiologie und begrifflicher Gliederung', Zeitschrift für romanische Philologie, 80: 51-69.

Laurén, Christer, and Heribert Picht. 1993. 'Vergleich Der Terminologischen Schulen'. In *Ausgewählte Texte zur Terminologie*. Wien: TermNet - International Network for Terminology.

Melnikow, G. 1988. *Systemology and Linguistic Aspects of Cybernetics*. Gordon & Breach Science Publishers Ltd.

Roche, Christophe. 2007. 'Le Terme et Le Concept: Fondements d'une Ontoterminologie'. In *TOTh 2007: « Terminologie & Ontologie: Théories et Applications »*. Annency: Institut Porphyre.

———. 2019. 'Tedi. Ontoterminology Editor. Manuel Utilisateur.' http://www.ontoterminology.com/.

Sowa, John F. 2000. 'Ontology, Metadata, and Semiotics'. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 1867: 55-81.

Wang, Mingyu. 2016. 'Toward the Meaning of Linguistic Signs: A Hierarchical Theory. Vol 2, N. 1.', Language and Semiotic Studies, 2 No. 1.

Wüster, Eugen. 1959. 'Das Worten der Welt, schaubildlich und terminologisch Dargestellt'. In *Terminologie Und Wissensordnung. Ausgewählte Schriften Aus Dem Gesamtwerk von Eugen Wüster*, edited by Heribert Picht and Klaus-Dirk Schmitz, 21-52. Wien: TermNet.

Résumé

Le but de cette étude est de comparer les modèles sémiotiques de Roche (2007) et Felber (1993). Tous deux traitent de la formalisation des connaissances, mais ils ont été développés à travers des influences théoriques, des traditions terminologiques et à des époques différentes. Alors que l'approche de Felber n'a jamais été opérationnalisée, le modèle sémiotique de Roche est à la base de l'éditeur d'ontoterminologie multilingue Tedi.

Les deux approches sont introduites puis les aspects du langage naturel, les concepts et la relation conceptuelle, le rôle de la logique et la représentation formalisée sont comparés. Pour englober tous ces aspects et se connecter à une application du monde réel, Tedi est utilisé comme une structure à laquelle les idées de Felber sont mises en correspondance. L'analyse montre que les approches de Roche et Felber diffèrent dans leur perspective sur le langage naturel, mais traitent à la fois des concepts et des relations de concepts, et utilisent la logique pour l'héritage des caractéristiques dans les structures de concepts hiérarchiques. Cela pourrait constituer un point de départ pour une analyse et une coopération plus approfondies.

Creating a termino-ontological resource for translators in the domain of viral infectious diseases

Alice Sanfilippo*

*Institute for Translation and Interpreting
University of Heidelberg
alice.snpp@gmail.com

Abstract. This paper describes the process of realizing a termino-ontological resource. The goal of this work is to propose a model for the creation of translation oriented terminological-ontological resources i.e., terminological databases supported by a formal ontology. The research question underlying the project is the following: can a formal ontology supporting a terminological resource be a tool to assist a translator in understanding the conceptual structure of the domain under analysis?

1. Introduction

The project presented in this paper consists of the realization of a termino-ontological resource: a multilingual terminological database for specialized translators supported by a formal ontology. Relying on good quality terminological resources, such as glossaries and specialized dictionaries, is in fact of key importance in the translation process and profoundly influences the quality of the final product. It is also common knowledge that the translator should have a good degree of knowledge of the topic of texts being translated. However, getting to know the domain to be translated is a task that requires a big amount of time and often traditional multilingual are usually limited to definitions and equivalent terms. Ontologies as means to organize knowledge appear to be a very useful tool that suits the purpose to introduce the translator to the domain of interest. At this purpose, in the last decades there has been a growing interest among terminology professionals for this kind of artifacts, and there are several studies that include the application of ontologies within the terminology activity.

Therefore, the idea behind the project is to create a translation-oriented terminological resource that, in addition to be a multilingual collection of terms, could help the translator to gain a coherent and unite picture of the domain under analysis.

In the following chapters, after describing the context that gave rise to this project, the methods used to create the resource will be described. In the description of the different phases followed during the process of realization of the resource a particular focus will be focus on the phase concerning the formalization of the ontology, since this is what makes a termino-ontological resource different than a traditional terminological database.

2. Theoretical background

This project builds its foundations in the acknowledgment of the importance of a conceptual system in terminological work, which has been stressed out since the first theory of terminology was developed by Eugene Wüster. Structuring and classifying knowledge is, in fact, fundamental in order to obtain a coherent image of the portion of world under analysis, i.e., the domain of interest. According to Wüster concepts are the starting point of every terminological work and the goal of the terminological activity is achieving a sharp distinction between the concepts (Wüster 1991, 1). Concepts are thus considered as the focus of the terminological activity, and the conceptual sphere is considered as independent from that of terms. Concepts are perceived as abstract entities that refer to objects in the real world and terms are only their linguistic representation. Fundamental according to the Austrian engineer in the analysis of terminology are the relationships existing between the concepts. Concepts and therefore terms are not to be considered independent units, but they must be analyzed in context, i.e., as parts of a conceptual system.

From this perspective, terminological work and the development of ontologies share in fact many commonalities. On the one hand, the goal of a terminographer is to collect all the terms concerning a specialized domain, to provide appropriate definitions and to identify the semantic relationships between them. On the other hand, the task of otologists is to organize the concepts of a given domain, identify their relationships, provide ambiguous definitions, encoded in a formal language, as well as allowing the exchange of unambiguous messages between software. It can therefore be inferred that both disciplines create mental organizations of the domains taken into analysis, and both aspire to unambiguous communication between the subjects through appropriate definitions (Spyns and De Bo 2004). For these reason in recent years different methodologies that foresee the application of ontologies in terminological work were developed. In particular relevant for the development of this project as theoretical backbone are: the Ontoterminology by Roche *et al*. (2009), the Termontography by Temmerman and Kerremans (2003) and Frame Based Terminology by Faber *et al*. (2005).

3. Realization of the termino-ontological database

3.1. Definition of the resource and documentation phase

The first phase in the realization of the termino-ontological database concerned the definition of the resources i.e., identification of the domain, scope and final users. The database is intended for native speakers of Italian who translate from German into Italian and is designed to support them in translating specialized texts in the field of viral infectious diseases. This is supported by a formal ontology that organizes the concepts in a structured way according to their location within the conceptual system and highlights the relationships between them. It aims to offer the translator, in addition to the translation of the terms, an overview of the domain and detailed information about its internal organization. Ideally, the final user has the possibility not only to search for terms, but also to access the ontological structure of the domain.

After defining the requirement that the resource has to fulfill a documentation phase was carried out. This allowed to gain as much information as possible about the domain of interest. During this phase, it was possible to identify the general concepts of the domain, the relationships between them, and the categorizations by topic to which they belong. At this stage is reasonable splitting the domain in smaller subdomains to gain a clearer and more organized overall view of the broader domain (Arntz, Picht, and Schmitz 2016). In addition, during the documentation stage, macro-thematic areas, to which each term belongs, were identified. These are: microbiology/virology, immunology, epidemiology, diagnostics, and prevention. This stage also involved the analysis of how the domain of infectious diseases or, more generally, the biomedical domain is covered in other terminological and knowledge organization systems. In a domain such as the biomedical one, in order to allow the correct interpretation of data and the standardization in the use of language, it is essential to organize knowledge and related terminology, both for epidemiological purpose, but also to ensure a proper management of the of the patient care process, both within a healthcare organization and among the different health care professionals. Exchanging such data is only possible when these are endowed with semantic interoperability, i.e. when computerized system are able to understand the meaning of the information stored in them (Duclos *et al*. 2014). For this reason, a broad range of knowledge organization systems exist in the biomedical domain. In particular the resources analyzed were the *International Classification of Diseases (ICD)*[1], *the Human Disease Ontology (DO) (*Schriml *et al.* 2019*)*, *SNOMED CT*[2] and the UMLS[3] (Unified Medical Language System). Each one presents different characteristics depending on the purpose they were built for. As will later be demonstrated, some of the elements of already existing resources were also used in this project.

3.2. Corpus compilation and terms extraction

In order to identify the terms to be included in the terminological database, an ad hoc corpus was compiled for the project. The texts that make up the corpus are in German, a choice motivated by the fact that the resource is designed for translators who translate Italian texts from German. German is therefore the source language of the bilingual terminological project. The corpus is made up of texts with different degrees of specialization: divulgative texts for a non-specialist audience, didactic texts addressed to health care disciplines students, specialized magazines, guidelines that are characterized by a more specialized use of the language. In addition to the source of the terms, the corpus along with the information gained in the documentation phase constitute the source of knowledge for the ontology. So, the concepts of the ontology are represented in this work by the German terms extracted from the corpus.

The corpus was uploaded and processed on the online platform Sketch Engine (Kilgarriff *et al*. 2014) which, among the various functions available, allows to extract terms automatically. Sketch Engine allows to extract both single words and multiword expressions. In the current work only substantives were included in the terminological database whereas the keywords extractor of Sketch Engine includes every part of speech. The obtained list of

[1] https://www.who.int/classifications/icd/en/
[2] https://browser.ihtsdotools.org/
[3] https://www.nlm.nih.gov/research/umls/index.html.

candidate terms was than reviewed, and the terms were manually validated, by confirming or rejecting each term.

Each of these has also been assigned a thematic macro-area of reference among those identified during the documentation phase. During this phase, both terms related to the main concepts identified in the first documentation phase as well as relevant terms that had not been taken into account in the previous phase were detected. After verifying their relevance, a further documentation phase was carried out. This consisted in searching for the term within the corpus, in order to identify its context of use and its position within the conceptual system.

3.3. Ontology creation

The formal ontology intended to support the terminological database was created through a middle-out approach (Uschold and Gruninger 1996). Therefore, after defining in the initial phase the fundamental and most important categories of the domain these were then generalized and/or specified. As previously mentioned, the method used involves acquiring knowledge of the domain of interest from a corpus, and after identifying the most relevant concepts, terms extracted from the corpus allowed the list of concepts to be expanded. The terms of the corpus correspond thus to concepts of the ontology. In this phase the relations among the concepts were also individuated. The ontology created consist of mostly hierarchical relationships, but also casual and functional relationships.

To develop the formal ontology in OWL[4] the open source ontology editor Protégé (Musen 2015) was used. The concepts of the ontology, and thus the classes in Protégé, are represented in the current work by the terms in German extracted from the corpus. Although they are part of two different dimensions of terminology, the linguistical one and the conceptual one.

The terms extracted from the corpus and the fundamental concepts identified during the first documentation phase were organized into hierarchies. In order to represent the domain of viral infectious disease of the central concept <Erreger> (pathogen) was used as starting point for the realization of the formal ontology. The scope of the ontology is in fact to formally represent the domain of the infectious disease of viral origin by considering an event

[4] OWL is a logic-based formal language for defining ontologies that allows knowledge to be expressed in such a way that it can be exploited by computer programs.

within it: the infection caused by the penetration of the pathogen into the host organism.

In order to recognize, within the list of candidate terms extracted from the corpus, those terms belonging to the conceptual category <Erreger>, all the compounds in German formed by the lexeme *Virus* were identified (e.g., *Varicella-zoster-Virus, Lyssaviren*)., which, along with the international abbreviations ending with the consonant *V* (e.g., HIV: Human Immunodeficiency Virus), appears to be the predominant form in the denomination of viruses.

All these terms have then been organized according to the type of nucleic acid (single or double-stranded RNA or DNA) that makes up their genome. Figure 1 shows the hierarchical classification of viruses in Protégé, where the different subclasses of <Virus> are subclasses of the class <Erreger>.

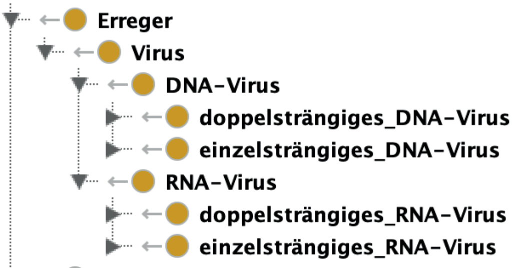
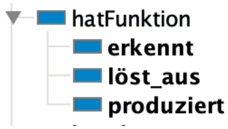

Erreger
Virus
DNA-Virus
doppelsträngiges_DNA-Virus
einzelsträngiges_DNA-Virus
RNA-Virus
doppelsträngiges_RNA-Virus
einzelsträngiges_RNA-Virus

FIG. 1 – *Viruses classification in Protégé*

However, identifying the subclasses of a class was not always that easy and straightforward. For this reason, and to further clarify the organization of the domain by grouping similar domain concepts into more general classes, some categories were borrowed from an existing medical terminological database: SNOMED CT. This is a collection of biomedical information organized in a systematic and computer-processable manner, providing codes, terms, synonyms, and definitions, used in clinical documentation. The purpose of the resource is to serve as a tool for sharing and reusing clinical information. SNOMED concepts can be combined to describe a more complex condition (IHTSDO 2018). SNOMED has a hierarchical conceptual organization consisting of 19 top-level concepts. This hierarchy has a high degree of granularity, which is considered too detailed given the scope of the current project which only aims at proposing a model. However, it seemed useful in order to give a more systematic organization to the ontology, once a concept has been identified, to include the corresponding superordinate concept in SNOMED. For instance, the concepts <Antikörper> and <Antigene> (antibody and antigen in SNOMED) in the SNOMED conceptual system are both subclasses of the class Immunologic Substance. Therefore, the class <ImmunologischeSubstanz> was also created in this ontology. In this way, a more precise organization has been achieved so that concepts belonging to the same categories are grouped together within the same more general super class.

The concepts of SNOMED were only used when considered necessary, and not systematically applied to all the classes of the ontology, i.e., when their application allowed a classification of the concepts that was as clear and structured as possible. Moreover, these concepts do not correspond to terms in the terminological database but are only intended to give the ontology a better organized structure. Within the ontology non-hierarchical relations can also be found, such as causal relationships, functional relationships, which relate a concept and its function. Like classes, relations are also organized in hierarchies. For example, given the top object property <hatFunktion> a series of relationships expressed in a different way, but represent as well functional relationships, were created – see the example in Figure 2.



hatFunktion
- erkennt
- löst_aus
- produziert

Fig. 2 – *Functional relationships hierarchy*

As already mentioned, a central event that takes place within the chosen domain is that of the infection, so ad hoc relations were created to represent it. In this way, the relation <hatÜbertragungsweg> (has routes of transmission) links the individuals of the class <Infektion> (infection) to those of the class <Übertragungsweg> (routes of transmission) and through an existential restriction in OWL all members of the class <Infektion> are related to an element of the class <Übertragungsweg> via the relation <hatÜbertragungsweg>. An existential restriction in OWL allows to describe all individuals of a class that are related to individuals of another class through a certain property (relationship). In general, all identified relationships are considered functional for the purposes of the project.

3.4. Combining linguistic and ontological data with LexO

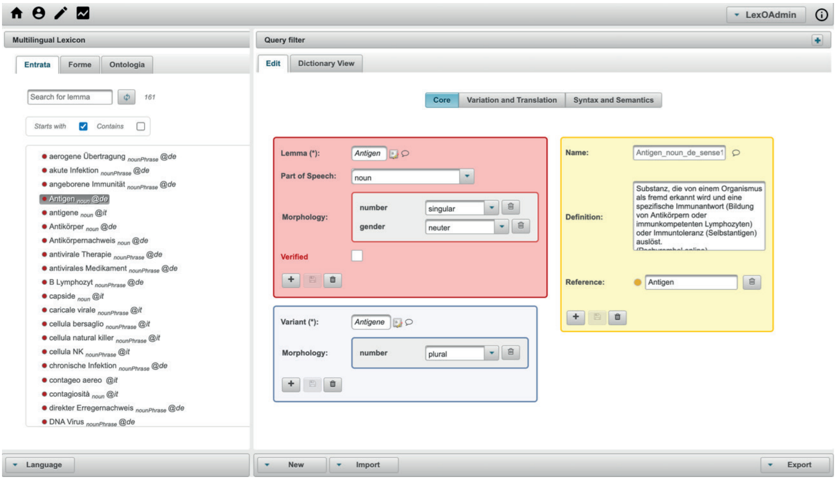
In order to combine the two aspects of the terminology in this project: the terminological and the ontological one, LexO (Bellandi *et al*. 2017; Bellandi, Giovannetti, and Piccini 2018) was used. LexO is a collaborative web editor developed by researchers at the Institute for Computational Linguistics "A. Zampolli" in Pisa of National Research Council of Italy[5]. The tool is designed to allow lexicographers and terminologists to create lexicographic and terminological resources that are linked to an ontology. LexO is based on the OntoLex-Lemon model (McCrae *et al*. 2017), the standard proposed by the W3C[6] to enrich ontologies with rich linguistic information and make them available for the Semantic Web and the Linguistic Linked Open Data. In the model each lexical entry is linked to an element (individual, class or relation) of the ontology.

LexO, as a tool designed for experts in the humanities who may not be familiar with the technical aspects of formal language, does not require special technical knowledge. In addition to representing resources for the Semantic Web, it offers a user-friendly consultation for the final user who has no knowledge of the formal language.

The interface of LexO (Figure 3) consists of two main sections. On the left side it is possible to navigate through the list of lemmas, forms or concepts of the ontology. In the case of a multilingual resources like the one in question, it is possible to filter these lists by language. In the central panel the editor and dictionary view are shown.

[5] http://www.ilc.cnr.it/

[6] The W3C (World Wide Web Consortium) is an international non-governmental organization that aims to promote the development of the full potential of the World Wide Web.



Search for lemma 161

Starts with ☑ Contains □

- aerogene Übertragung nounPhrase @de
- akute Infektion nounPhrase @de
- angeborene Immunität nounPhrase @de
- Antigen noun @de
- antigene noun @it
- Antikörper noun @de
- Antikörpernachweis noun @de
- antivirale Therapie nounPhrase @de
- antivirales Medikament nounPhrase @de
- B Lymphozyt nounPhrase @de
- capside noun @it
- caricale virale nounPhrase @it
- cellula bersaglio nounPhrase @it
- cellula natural killer nounPhrase @it
- cellula NK nounPhrase @it
- chronische Infektion nounPhrase @de
- contageo aereo @it
- contagiosità noun @it
- direkter Erregernachweis nounPhrase @de
- DNA Virus nounPhrase @de

Lemma (*): Antigen
Part of Speech: noun
Morphology: number singular; gender neuter
Verified □

Variant (*): Antigene
Morphology: number plural

Name: Antigen_noun_de_sense1
Definition: Substanz, die von einem Organismus als fremd erkannt wird und eine spezifische Immunantwort (Bildung von Antikörpern oder immunkompetenten Lymphozyten) oder Immuntoleranz (Selbstantigen) auslöst.
Reference: Antigen

Fig. 3 – *LexO Interface*

LexO in accordance with the Onto-Lex Lemon model consists of three modules: Core Module, Variation and Translation Module and Syntax and Semantics Module.

The ontology previously created with Protégé was uploaded to LexO in order to associate the concepts of the ontology with terms to which detailed linguistic information is added. At the moment in LexO (which is still a software under development) it is only possible to display the class hierarchy in the section dedicated to the ontology, the section concerning the relations has yet to be implemented. In order to create a bilingual resource, such as the one in question, a lexicon has been created whose lexical entries in Italian and German will be, according to the OntoLex-Lemon model, instances of the LexicalEntry class.

As shown in Figure 3, linguistic information is added to each lexical entry. In the red box the canonical form of each term is displayed to this the morphological traits of gender and number have been added. In the blue box another LexicalForm of the lemma can be added, in this case the plural variant.

In the yellow box through LexicalSense[7] the meaning of each lexical entry is specified through a definition and the linking of this to an ontological entity (indicated by Reference). Each lexical entry can have different lexical senses. This is the case of polysemantic words which have as much lexical sense as the meaning of the word. In the Variation and Translation Module it is possible to express lexical relationships between terms, such as for example a relationship between the extended form of a term and its abbreviation or acronym. In this module, it is also possible to include the translation of a term into a different language, in particular by linking two equivalent terms in the two languages. According to OntoLex-Lemon, this type of equivalence can be expressed at three different levels[8]:

- at the ontological level: when two entries in different languages denote the same ontology entity;
- at the lexical level: when the lexical entries might not denote exactly the same concept, but their lexical meanings (senses) can be exchanged for each other in most contexts;
- via the relation translatableAs: when a lexical entry can be translated into another entry only in some contexts, specifying in which context and the exact lexical senses under which this translation is valid.

In the case of the project in question, the relationship of equivalence between two terms has been expressed at two levels: ontological and lexical level. Thus, two equivalent terms are linked by a sense relationship, and, in addition, they share the same ontological reference, in order to make explicit that both terms are the linguistic representation of the same concept. See the example in Figure 4 where the two lexical entries *Antigen* and *antigene* have the same ontological reference and each one is the translation of the other.

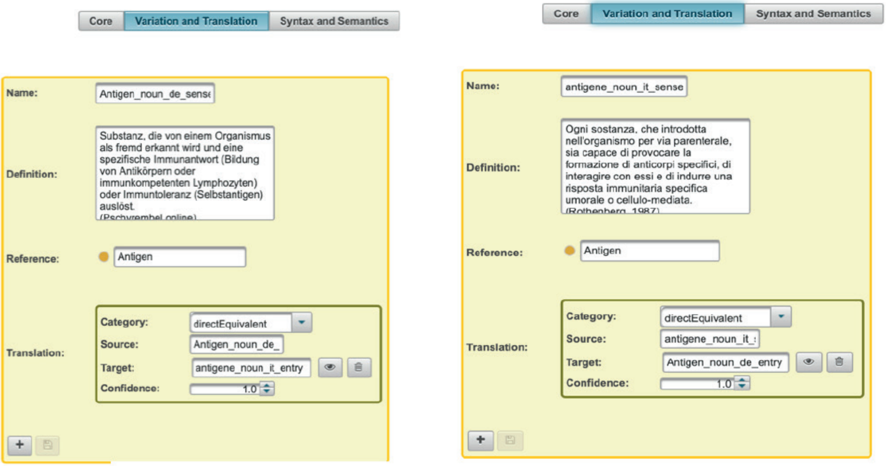
In addition, LexO offers the possibility to choose between three different kinds of equivalents, which according to the vocabulary[910,] used are:

[7] In the Ontolex-Lemon model LexicalSense is the linking between a lexical entry and the ontological element.

[8] https://www.w3.org/2016/05/ontolex.

[9] https://github.com/andreabellandi/LexO-lite.

[10] http://linguistic.linkeddata.es/def/translation-categories.



Name: Antigen_noun_de_sense

Definition: Substanz, die von einem Organismus als fremd erkannt wird und eine spezifische Immunantwort (Bildung von Antikörpern oder immunkompetenten Lymphozyten) oder Immuntoleranz (Selbstantigen) auslöst. (Pschyrembel online)

Reference: ● Antigen

Translation:
Category: directEquivalent
Source: Antigen_noun_de_
Target: antigene_noun_it_entry
Confidence: 1.0

Name: antigene_noun_it_sense

Definition: Ogni sostanza, che introdotta nell'organismo per via parenterale, sia capace di provocare la formazione di anticorpi specifici, di interagire con essi e di indurre una risposta immunitaria specifica umorale o cellulo-mediata. (Rothenberg, 1987)

Reference: ● Antigen

Translation:
Category: directEquivalent
Source: antigene_noun_it_
Target: Antigen_noun_de_entry
Confidence: 1.0

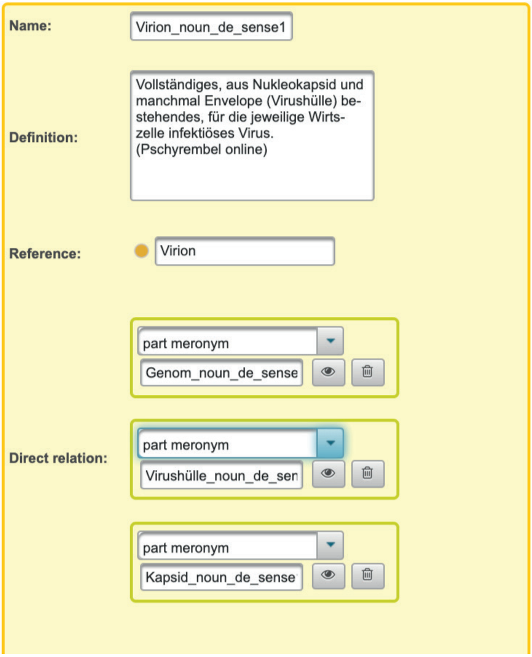
FIG. 4 – *Representation of equivalent terms in LexO*

- direct equivalent: when the two terms describe entities that are semantically equivalent and refer to entities that exist in both languages and cultures;
- cultural equivalent: when two terms describe entities that are not equivalent at the semantic level, but at the pragmatic level, so that they describe similar situations in different cultures and languages;
- lexical equivalent: used for those terms in different languages that refer to the same entity, but one of them express the original term using words from the target language.

In the current project equivalents are mostly direct.

In the Variation and Translation module, it is also possible to link lexical entries through sense relations. This function is particularly useful in the realization of the resource in question, since one of its objectives is to highlight the relationships existing between concepts and therefore between terms. It is therefore possible to establish sense relations between the lexical entries, such as those of the ontology. For instance, the partitive relationship that in the ontology is represented by the object property <hatBestandteil> (has component) is realized at the terminological level in LexO through holonymy and meronymy relations. Figure 5 shows how the term Virion (holonym) is linked through a partitive relation to its components, i.e., to the elements constituting the concept <Virion>. The lexical entries *Genom*, *Virushülle* and *Kapsid* are thus meronyms of *Virion*.

In the same way the corresponding lexical entries in Italian are linked via the same type of relationship.



Name: Virion_noun_de_sense1

Definition: Vollständiges, aus Nukleokapsid und manchmal Envelope (Virushülle) bestehendes, für die jeweilige Wirtszelle infektiöses Virus. (Pschyrembel online)

Reference: ● Virion

Direct relation:

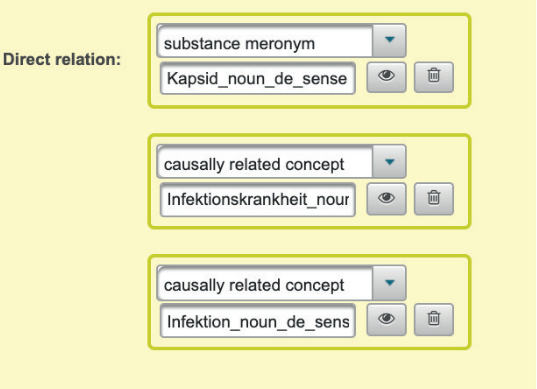
part meronym
Genom_noun_de_sense

part meronym
Virushülle_noun_de_sen

part meronym
Kapsid_noun_de_sense

FIG. 5 – *Partitive relationship in LexO*

It is also possible to connect lexical entries through the sense relationship "causally related concepts" (see Figure 6), establishing a causal relationship between the lexical entries analogously with their ontological reference. A casually related concept is defined by Lexinfo, the vocabulary used by OntoLex Lemon and thus by LexO, as "a concept that is related to another concept by virtue of the fact that it plays a causative role with respect to that concept”[11].



Direct relation:

substance meronym
Kapsid_noun_de_sense

causally related concept
Infektionskrankheit_nour

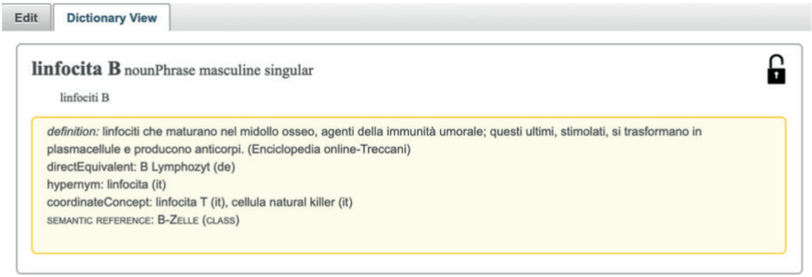
causally related concept
Infektion_noun_de_sens

Fig. 6 – *Causal relationship in LexO*

Finally, the end user will be able to view all the linguistic and semantic information included in the different modules through the Dictionary View. As shown in Figure 7, in the dictionary view the hyperonym of the selected lexical entry and the co-ordinated concepts are shown. A co-ordinated concept is according to Lexinfo “a subordinate concept having the same nearest superordinate concept and same criterion of subdivision as some other concept in a given concept system”[12]. By browsing through the concepts of the ontology and selecting one of them, it is also possible to display the lexical entry(s) associated with them and the hierarchies of the concepts (Figure 8).

[11] https://lexinfo.net/ontology/3.0/lexinfo.

[12] *Ibidem*.

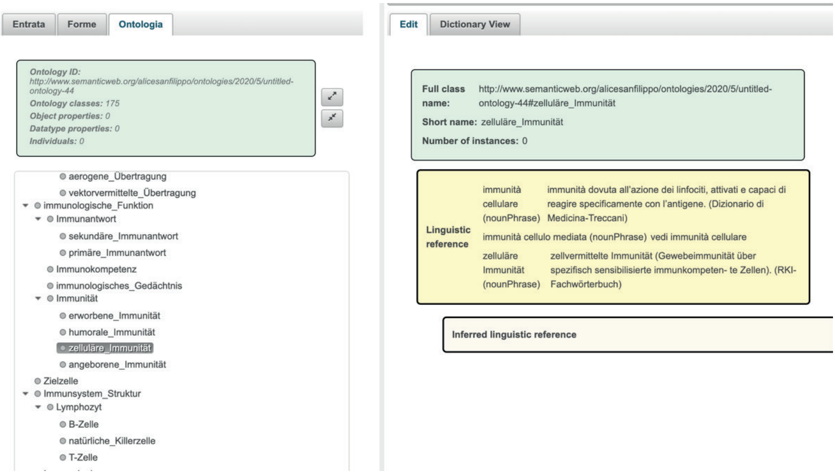


linfocita B nounPhrase masculine singular

linfociti B

definition: linfociti che maturano nel midollo osseo, agenti della immunità umorale; questi ultimi, stimolati, si trasformano in plasmacellule e producono anticorpi. (Enciclopedia online-Treccani)
directEquivalent: B Lymphozyt (de)
hypernym: linfocita (it)
coordinateConcept: linfocita T (it), cellula natural killer (it)
SEMANTIC REFERENCE: B-ZELLE (CLASS)

FIG. 7 – *Dictionary View in LexO*



Ontology ID: *http://www.semanticweb.org/alicesanfilippo/ontologies/2020/5/untitled-ontology-44*
Ontology classes: 175
Object properties: 0
Datatype properties: 0
Individuals: 0

- aerogene_Übertragung
- vektorvermittelte_Übertragung
- immunologische_Funktion
  - Immunantwort
    - sekundäre_Immunantwort
    - primäre_Immunantwort
  - Immunokompetenz
  - immunologisches_Gedächtnis
  - Immunität
    - erworbene_Immunität
    - humorale_Immunität
    - zelluläre_Immunität
    - angeborene_Immunität
- Zielzelle
- Immunsystem_Struktur
  - Lymphozyt
    - B-Zelle
    - natürliche_Killerzelle
    - T-Zelle

Full class name: http://www.semanticweb.org/alicesanfilippo/ontologies/2020/5/untitled-ontology-44#zelluläre_Immunität
Short name: zelluläre_Immunität
Number of instances: 0

Linguistic reference
immunità cellulare (nounPhrase) immunità dovuta all'azione dei linfociti, attivati e capaci di reagire specificamente con l'antigene. (Dizionario di Medicina-Treccani)
immunità cellulo mediata (nounPhrase) vedi immunità cellulare
zelluläre Immunität (nounPhrase) zellvermittelte Immunität (Gewebeimmunität über spezifisch sensibilisierte immunkompeten- te Zellen). (RKI-Fachwörterbuch)

Inferred linguistic reference

FIG. 8 – *Browsing through the ontology hierarchy in LexO*

The result is a terminological database that combines in itself a semasiological approach, where lexical entries are listed alphabetically, and an onomasiological approach, where terms are grouped according to the concept they denote, and the relationships between them.

The resources created with LexO are created according the the Semantic Web standards, OWL and OntoLex-Lemon, and can therefore be exported in these formats and shared in the Semantic Web.

4. Conclusion

The goal of this project was to create an experimental model to verify if the method of combining terminologies and ontologies could be functional in supporting technical translators in getting to know the domain they translate. Given this premise, it is believed that the resource created, also with the help of LexO, can be functional and that this same strategy of action can be applied to different domains. However, in the case of other domains it is useful to verify the existence of knowledge organization systems for the domains of interest, especially if one does not have the support of domain experts.

The creation of an ontology is an activity that presents quite a few difficulties, first of all because the conceptualization of a specialized domain is not easy for a non-expert in the field. In addition, the domain under study appears particularly complex because it involves a number of processes that can be fully understood only after an in-depth study. However, the wide availability of resources in the biomedical domain aimed at collecting and organizing knowledge has facilitated the creation of the ontology.

LexO proved to be a suitable tool for the purpose of the project. In fact, thanks to this software it was possible to create a database that separated the two levels of terminology: the linguistic and the conceptual one. However, there are some elements of the formal ontology created within the project that cannot be visualized in LexO because, being a software still in development phase, it has not yet implemented this functionality. However, it was possible to highlight different types of relationships between concepts and therefore between terms, thanks to the relationship between senses. A functionality that could support the translator in understanding the domain and provide an overall structure of its conceptual system is the possibility of navigating between the classes of the ontology and then searching for terms based on the concept with which they are associated. This function is however not available to the end user in the current version of LexO. In addition, a feature that could serve the scope of this project is a possibility to display a graphical representation of the concept system of the domain under analysis.

References

Arntz, Reiner, Heribert Picht, and Klaus-Dirk Schmitz. 2016. Einführung in die Terminologiearbeit. Georg Olms Verlag.

Bellandi, Andrea, Emiliano Giovannetti, and Silvia Piccini. 2018. ‘Collaborative Editing of Lexical and Terminoontological Resources: A Quick Introduction to LexO’. In The XVIII EURALEX International Congress. Lexicography in Global Contexts, 23-27.

Bellandi, Andrea, Emiliano Giovannetti, Silvia Piccini, and Anja Weingart. 2017. ‘Developing LexO: A Collaborative Editor of Multilingual Lexica and Termino-Ontological Resources in the Humanities’. In Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017). Montpellier, France: Association for Computational Linguistics.

Duclos, C., A. Burgun, J. B. Lamy, P. Landais, J. M. Rodrigues, L. Soualmia, and P. Zweigenbaum. 2014. ‘Medical Vocabulary, Terminological Resources and Information Coding in the Health Domain’. In Medical Informatics, e-Health: Fundamentals and Applications, edited by Alain Venot, Anita Burgun, and Catherine Quantin, 11-41. Health Informatics. Paris: Springer.

Faber Benítez, Pamela, Carlos Márquez Linares, and Miguel Vega Expósito. 2005. ‘Framing Terminology: A Process-Oriented Approach’. Meta 50 (4).

IHTSDO. 2018. ‘SNOMED CT Starter Guide - SNOMED CT Starter Guide - SNOMED Confluence’.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. ‘The Sketch Engine: Ten Years On’. Lexicography 1 (1): 7-36.

McCrae, John P., J. Gil, J. Gràcia, Paul Bitelaar, and P. Cimiano. 2017. ‘The OntoLex-Lemon Model: Development and Applications’. Undefined.

Musen, Mark A. 2015. ‘The Protégé Project: A Look Back and a Look Forward’. AI Matters 1 (4): 4-12.

Roche, Christophe, Marie Calberg-Challot, Luc Damas, and Philippe Rouard. 2009. ‘Ontoterminology: A New Paradigm for Terminology’. In International Conference on Knowledge Engineering and Ontology Development, 321-26. Madeira, Portugal.

Schriml, Lynn M., Elvira Mitraka, James Munro, Becky Tauber, Mike Schor, Lance Nickle, Victor Felix, *et al*. 2019. ‘Human Disease Ontology 2018 Update: Classification, Content and Workflow Expansion’. Nucleic Acids Research 47 (D1): D955-62.

Spyns, Peter, and Jan De Bo. 2004. 'Ontologies: A Revamped Cross-Disciplinary Buzzword or a Truly Promising Interdisciplinary Research Topic?' Linguistica Antverpiensia, New Series–Themes in Translation Studies, no. 3.

Temmerman, R., and K. Kerremans. 2003. 'Termontography: Ontology Building and the Sociocognitive Approach to Terminology Description'.

Uschold, Mike, and Michael Gruninger. 1996. 'Ontologies: Principles, Methods and Applications'. *Knowledge Engineering Review* 11: 93-136.

Wüster, Eugen. 1991. *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. 3. Aufl., mit Einem Vorw. von Richard Baum. Abhandlungen zur Sprache und Literatur: ASL. Bonn: Romanistischer Verlag.

Résumé

Le projet présenté dans cet article consiste dans la réalisation d'une ressource termino-ontologique: une base de données terminologiques multilingue destinée aux traducteurs spécialisés dans le domaine des maladies infectieuses, soutenue par une ontologie formelle. S'appuyer sur des ressources terminologiques de bonne qualité est en effet d'une importance capitale dans le processus de traduction et d'influence dans la qualité du produit final. Il est également bien connu que le traducteur doit avoir un bonne connaissance du sujet des textes qu'il traduit. Cependant, apprendre à connaître le domaine à traduire est une tâche qui demande beaucoup de temps au traducteur et souvent les ressources multilingues traditionnelles manquent de ce type d'information et se limitent généralement aux définitions et aux termes équivalents. Par conséquent, l'idée à l'origine du projet était de créer une ressource terminologique qui, en plus d'être une collection multilingue de termes, pourrait aider le traducteur à obtenir une image cohérente et unie du domaine analysé.

Crisis in troubled ancient times: ontological modelling of textual evidence from Greek historians

Maria Papadopoulou*, Eleni-Melina Tamiolaki**, Christophe Roche*

*University Savoie Mont-Blanc, Condillac-LISTIC, France & Liaocheng University, KETRC, China, maria.papadopoulou@univ-savoie.fr, christophe.roche@univ-savoie.fr
** University of Crete, Department of Philology & IMS/FORTH, tamiolaki@uoc.gr

Abstract. This article presents work-in-progress towards the construction of a model of crises for ancient Greek historiography and a terminology of crises with equivalents in Greek (ancient and modern) and English. In order to solve the problem of linking and aligning these three vocabularies as well as defining their meaning, we rely on a common conceptualization of crises. This conceptualization is represented as an ontology built in a computer readable form compliant with the W3C standards for opening and linking data on the Semantic Web. By means of the temporal and causal relational information not explicitly in any of historiographical resources, the ontology facilitates reasoning on and across documents, revealing relationships between events to answer complex questions.

> *Pourquoi tirons-nous tant de jouissance d'être si différents non seulement des autres mais de notre propre passé? Quel psychologue assez fin expliquera cette délectation morose à être en crise perpétuelle et à finir l'histoire?*
> *Why do we get so much pleasure out of being so different not only from others but from our own past? What psychologist will be subtle enough to explain our morose delight in being in perpetual crisis and in putting an end to history?*
> Bruno Latour, 1993, *We have never been modern*, Ch. 1 "Crisis" [Transl. C. Porter]

1. Introduction

Crises can leave an indelible mark on history. Crises are commonly defined as turning points in a sequence of events that can determine future events. Crisis definitions and the terminology around crises are, at heart, political issues [McConnell 2022]. One cannot deal with the long and complicated history of a nation or of an entire period without referring to the great crises and it is the salient crises that are most likely to form the political memory of a people [Verba 1965, 55].

This article presents research done towards building a model for the computational representation of crisis events using an ontology encoded in the Web Ontology Language -OWL [OWL 2004]. The ontology aims principally at the semantic annotation that enables the semantic querying of ancient Greek historiographical texts and the definition of the terminology of crisis in ancient Greek times. This work is done within the framework of the *Leaders and Crisis Management in Ancient Literature. A Comparative Approach* (LACRIMALit) project[1]. Scholarship so far has studied the phenomenon of crisis mainly as times of intense difficulty or danger linked with a specific historical period and sometimes even with one single aspect (financial, social, political, cultural, etc.) of human activity. Most of these studies do not provide a thorough examination of the term "crisis", they do not analyse the elements which form a crisis narrative and they do not attempt a comparison between the various crisis narratives. The LACRIMALit project aims to fill this gap by investigating contemporary theories on the notion of crisis [Engels 2012; Hermann 2011], applying them to the study of ancient texts and proposing analogies with contemporary debates, when needed, with caution and while acknowledging the different historical and political contexts.

As part of the LACRIMALit project, the LACRIMALit ontoterminology of crisis is built taking into account different sources reporting crisis incidents in Greco-Roman antiquity. As structured descriptions of events are either missing or not marked up explicitly, the LACRIMALit ontoterminology of

[1] The *Leaders and Crisis Management in Ancient Historiography. A Comparative Approach* (LACRIMALit) is a project hosted at the Institute for Mediterranean Studies/Foundation for Research and Technology (IMS-FORTH), 2022-2025. The project has received funding by the Hellenic Foundation for Research and Innovation (H.F.R.I). Principal Investigator: Eleni-Melina Tamiolaki, University of Crete, Greece. The acronym is a pun whose first component (*lacrima* = tear, in Latin) suggests a link between tears and the notion of crisis as a turning point that can reduce one to tears. Project website: https://ims.forth.gr/en/project/view?id=219.

crisis aims to narrow this existing gap. Also, the LACRIMALit structured vocabulary of crisis-related terms in Greek (in both its ancient and modern varieties) and English is usable by both humans (via an electronic dictionary interface) and machines (via the W3C standard languages). It thus supports consistency in the terms used, open and linked data discovery across multiple sources [Heath and Bizer 2011], automated reasoning upon the modelled data, and, finally, semantic annotation of the relevant textual sources with the terms linked to the ontology. The overall aim of ontology-based semantic annotation of these ancient texts is the promotion of new affordances in their reading online, now possible only via the hyperlinked e-texts openly accessible by the Perseus Digital Library [Crane 2012]. The ontoterminological resources to be created by the completion of the project will be of use to classical scholars as well as comparative politics experts making at their disposal a computerised model of crisis-related shared terms in Linked Open Data formats, i.e., tagged in a consistent and interoperable manner increasing their findability and accessibility.

2. The LACRIMALit project

The LACRIMALit ontology project is part of a broader research which aims to offer a comprehensive study of the phenomenon of leadership and crisis management in Greco-Roman Antiquity focusing on four genres of ancient Greek literature (historiography, rhetoric, comedy and biography). It also relies on theoretical texts dealing with the issue of leadership in Antiquity, such as Xenophon's *Hieron*, Isocrates' *Nicocles* and *To Nicocles*, the first two books of Aristotle's *Politics*, the sixth book of Polybius' Histories and Cicero's *De re publica*. Our work on the ontology of crisis will start with studying crisis in ancient historiography.

3. Modelling Events in Digital Humanities

Although still in an unsystematic and uncoordinated fashion, ontologies are more and more used in the Digital Humanities [Jansen, 2019][2]. As Cybulska & Vossen 2010 aptly remark, the ability to automatically determine relations between historical events and their sub-events over textual data, based on the relations between event participants, time markers and locations,

[2] For a curated list of ontologies in Digital Humanities see https://github.com/CLARIAH/awesome-humanities-ontologies.

have important repercussions for the design of historical information retrieval systems. Ontologies, however, are mostly manually constructed.

The LACRIMALit model for crisis in ancient historiography is an event centric model. Several event models have been published over the past years in different humanities domains, notably in the domains of history and cultural heritage. In what follows we include the definitions of events in event-centric models most relevant to the LACRIMALit model, starting from the ontological definition of a recent paper [Guarino *et al.* 2022]. The paper asks two central questions in regard to the development of event-centric ontologies, the first focusing on metaphysics, the other on semantics: ‘What are events?’ and ‘What is the referential mechanism that is in play when we describe an event?’ The short answer to the first question is that the simplest events are qualitative changes cognitively constructed. The short answer to the second question is that the notion of event is intimately connected to that of context, i.e., describing an event means not just saying what happened, but also specifying how it happened, by specifying details that often involve the *context* in which the event occurred.

From an ontological, formalizable, viewpoint, an event is anything that has occurred in a certain time and environment where some actors could take part and show some action features. In contrast to “objects” or “continuants’, which take up space, *are in* time and persist through time by being wholly present at every time at which they exist, “events” or “occurrents” (i.e., entities that occur or happen), also referred to as “perdurants”, or “processes”, are four-dimensional: they *take up* time and persist through time by having different “stages” [Maienborn 2011; Galton 2012; Arp *et al.* 2015, 87; Rodrigues and Abel 2019; Casati and Varzi 2020]. In event calculus terms events are “fluents” (i.e., statements representing properties that vary over time, e.g., the number of a person’s children at different times).

Event-centred modelling captures the dynamic aspects of a domain. In addition, events provide a natural way to explicate complicated relations between people, places, actions and objects [van Hage *et al.* 2011]. Events are central elements in the representation of data from various fields such as history and cultural heritage. Doerr and Kritsotaki 2006 propose to see events as meetings, that are, in turn, interactions of participants which bring about changes of state. Corda *et al.* 2011 identify events in the domain of the history of science as situated occurrences incorporating complex and rich information about the subject of the event (who), the object (what), the time (when), the place (where), the cause(s) and effect(s) (why).

3.1. CIDOC-CRM

The literature on modelling events is vast[3]. Different event models provide different definitions of events. Here we present models of interest for the domain of LACRIMALit ontology. Of the models included below, at its present state, LACRIMALit is only aligned to CIDOC-CRM.

The most important model for the domain of digital humanities and cultural heritage is CIDOC-CRM, therefore LACRIMALit is aligned to CIDOC. The CIDOC Conceptual Reference Model (CRM), [ISO 21127:2014, first edition ISO 21127:2006] is a high-level, event-centric, formal ontology of things and events happening in spacetime. In CIDOC the E5_Event class comprises "distinct, delimited and coherent processes and interactions of a material nature, in cultural, social or physical systems, involving and affecting instances of E77_Persistent Item in a way characteristic of the kind of process. Typical examples are meetings, births, deaths, actions of decision taking, making or inventing things, but also more complex and extended ones such as conferences, elections, building of a castle, or battles" (CIDOC Version 7.7.1, April 2021)[4]. Among the different subclasses of E5_Event, the E7 Activity comprises "actions intentionally carried out by instances of E39_Actor that result in changes of state in the cultural, social, or physical systems documented. This notion includes complex, composite and long-lasting actions such as the building of a settlement or a war, as well as simple, short-lived actions such as the opening of a door" (CIDOC Version 7.7.1, April 2021).

[3] For a comparative description of five existing event models, see [Astrova et al. 2014]. Indicatively, the ABC Ontology for digital libraries, whose purpose was to facilitate interoperability between metadata vocabularies from different domains. [Lagoze and Hunter 2001]; the Event ontology [Raimond and Abdallah 2006] is an event-centric model for the domain of Music, which defines events as arbitrary classifications of space/time regions by a cognitive agent that may have participating agents, passive factors, products, and a location in space/time. Event-Model-F [Scherp et al. 2009] designed to facilitate interoperability in distributed event-based systems. The model is based on the foundational ontology DOLCE+DnS Ultralite (DUL) (DnS = Descriptions and Situations). It provides comprehensive support for the representation of time and space, objects and persons, as well as the mereological, causal and associative relations between events. Event-Model-F provides a means for event composition, modelling event causality and event correlation, and representing different interpretations of the same event, and can be easily extended by domain specific ontologies.

[4] [Bekiari et al. 2021]. Accessible online: https://cidoc-crm.org/version/version-7.1.1.

LACRIMALit events fall under class E7_Activity, a subclass of E5 Event, as LACRIMALit extends CIDOC[5].

3.2. LODE (Linking Open Descriptions of Events)

LODE *(Linking Open Descriptions of Events)* [Shaw *et al*. 2009][6] is an ontology for publishing descriptions of historical events as Linked Data, and for mapping between other event-related vocabularies based on what happened, where something happened, when it happened, and who was involved. These "factual" relations within and among events are constructed to generate representations of "intersubjective consensus of reality" not necessarily associated with a particular perspective or interpretation of one principal class (Event) and seven properties that refer to the happening of the event. The "Event" class is defined as "something that happened", which has temporal and spatial boundaries, thus enabling statements correlated to people, places or things. By this definition some particular event does not necessarily involve state change. Additionally, events are not differentiated from processes or states.

LODE defines one class Event class "Something that happened," as might be reported in a news article or explained by a historian. LODE defines two properties for location where an event happened atPlace, for a named or relatively specified place, and inSpace, for an abstract region of space, e.g., a geospatial point or region. Also, LODE has two properties of time atTime for abstract instants or intervals of time and circa property for precise intervals of time, such as calendar dates and clock times. Finally, LODE defines two properties for an agent or object (physical, social, or mental), respectively.

3.3. SEM (Simple Event Model)

The Simple Event Model (SEM) was created to model events in various domains, without making assumptions about the domain-specific vocabularies used. It is presented by virtue of two use cases: historic events and events in the maritime safety domain [van Hage *et al*. 2011][7]. Events, according to SEM, describe everything that happens, including fictional events. SEM classes are

[5] [Casati and Varzi 2020: 2.1] distinguish "events" into "activities", "accomplishments", "achievements", and "states".

[6] Available online: LODE: An ontology for Linking Open Descriptions of Events (linkedevents.org).

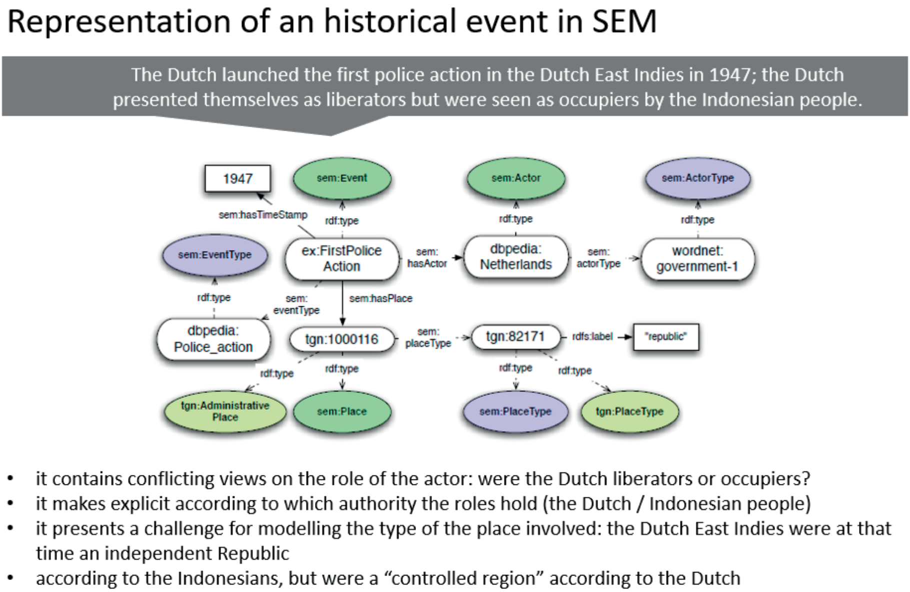
[7] Available online: https://semanticweb.cs.vu.nl/2009/11/sem/

divided into three categories: Core classes, Types, and Constraints. There are four core classes: sem:Event (what happens), sem:Actor (who or what participated), sem:Place (where), sem:Time (when). The SEM Type class contains all types of Core instances. These can be either individuals or classes themselves. This class is meant to be extended for each application domain. There are three kinds of Constraints: Role, Temporary and View. sem:Role describes the role that an individual of a class is playing in the context of a specific event. Roles can be specified for all Core individuals. The SEM Constraint class contains instances of properties that have a constrained (i.e., not universal) validity. This includes time-dependent validity (Temporary), validity in the guise of a specific role (Role), or validity according to a given Authority (View). Each core class has an associated sem:Type class, which contains resources that indicate the type of a core individual. Individuals and their types are usually borrowed from other vocabularies, e.g., Getty Thesaurus of Geographical Names (TGN)[8].

SEM's properties are divided in three kinds: sem:eventProperty, sem:type properties and a few other properties like sem:accordingTo and sem:hasTimeStamp's subproperties. The sem:eventProperty relates sem:Events to other individuals. A sem:type relates individuals of the sem:Core class to individuals of sem:Type. There are subproperties of sem:type for each of the separate core classes to facilitate querying. To represent opinions sem:accordingTo relates a sem:View to a sem:Authority. In terms of Time, the sem:hasTimeStamp property is for single time values, while for time intervals SEM has two properties (sem:hasBeginTimeStamp and sem:hasEndTimeStamp), and for uncertain time intervals, SEM has four properties (sem:hasEarliestBeginTimeStamp, sem:hasLatestBeginTimeStamp, sem:hasEarliestEndTimeStamp, and sem:hasLatestEndTimeStamp).

A historical event that occurred in 1947 Indonesia (Dutch East Indies, at the time) - the first police action in the Dutch East Indies in 1947 by the Dutch, who presented themselves as liberators, but were seen as occupiers by the Indonesian people - as represented in SEM is shown in Figure 1:

[8] Available online: https://www.getty.edu/research/tools/vocabularies/tgn/



Representation of an historical event in SEM

The Dutch launched the first police action in the Dutch East Indies in 1947; the Dutch presented themselves as liberators but were seen as occupiers by the Indonesian people.

- it contains conflicting views on the role of the actor: were the Dutch liberators or occupiers?
- it makes explicit according to which authority the roles hold (the Dutch / Indonesian people)
- it presents a challenge for modelling the type of the place involved: the Dutch East Indies were at that time an independent Republic
- according to the Indonesians, but were a “controlled region” according to the Dutch

Fig. 1 – *Representation of an historical event in SEM. Adapted from [van Hage et al. 2011, fig. 3]*

3.4. REO (Rich Event Ontology)

The goal of the Rich Event Ontology (REO) is to provide a unified representation of events with a rich structure of event concepts that connects varying levels of event specificity, relates events to their key objects and participants, and encodes the temporal and causal relationships between events. REO aims to bridge the gap between spatiotemporal ontological approaches to representing events and the representations stemming from semantic role labelling (SRL) resources. Unifying NLP resources, such as the FrameNet, VerbNet, the Rich Event Ontology (REO) [Brown *et al.* 2021] marries ontology with lexical resources (corpora) and serves as a shared hub for the disparate annotation schemas.

REO supports mapping between specific event types of different resources and enables the merging of associated annotated corpora and expanding sets of related event triggers. By adding temporal and causal relational information that does not exist in these resources, REO facilitates reasoning on and across documents, revealing relationships between events that come together in temporal and causal chains [Chiarcos *et al*. 2020:15].

To capture some of the rich and complex relations between events or between events and objects, REO includes temporal and causal relations extended from the Richer Event Description (RED) project [Ikuta *et al*., 2014; O'Gorman *et al*., 2016], such as the *hasPrecondition*, *hasCause*, *hasResult*, and *hasSubevent* relations. The RED project aims to annotate text with mentions of events and entities, with the goal of representing the temporal and causal relationships between those events in such a way that an accurate timeline of events could be automatically constructed.

4. Competency Questions

Competency Questions [Ren *et al*., 2014] play an important role in the lifecycle of engineering an ontology. Competency questions represent the requirements that an ontology has to fulfil.

At its present state, the LACRIMALit ontology is "competent" to answer the following set of competency questions (CQ):

- CQ1: What are the different types of political crises?
- CQ2: When did a crisis occur?
- CQ3: Where did a crisis take place?
- CQ4: Who are the protagonists of a crisis (e.g., Peloponnese war)?
- CQ5: What are the relevant passages in the primary sources of an event?
- CQ6: What are the relevant terms denoting crises (military, political etc.)?

5. Modelling Crises

Defining a crisis is quite complex because of the interdisciplinary nature of the concept. According to the Oxford Dictionary[9] it is defined as a point in time: 1. A time of intense difficulty or danger. 1.1 A time when a difficult or

[9] Oxford English Dictionary (online) s.v. crisis https://www.lexico.com/definition/crisis.

important decision must be made. 1.2 The turning point of a disease when an important change takes place, indicating either recovery or death[10].

Historically, increasing numbers of crises/disasters, natural and human-made, have demonstrated the importance of crisis management. The success of crisis management largely depends on finding, assembling, and successfully integrating related information in order to inform both the decision-making/response stage, as well as planning the preparedness/planning stage. Also, the degree of predictability of a crisis event is crucial: a crisis is predictable, if place, time or in particular the manner of its occurrence are knowable to at least one concerned party and if the probability of occurrence is not negligible.

Despite extensive relevant work on the importance of building a typology of crises in recent decades [Coombs 1998, Bringmann 2003, Franchet d'Espèrey *et al.* 2003, Gundel 2005, Rousseaux & Lhoste 2010, Angiolillo *et al.* 2015, Björck 2016], no such satisfactory typology exists. As [Björck 2016] succinctly puts it: "A typology is valuable because it simplifies and structures complexity, helps to organise the collection of information, provides diagnostic insights [Burnett, 1998] and is a first step to contain a crisis [Gundel, 2005]." According to Gundel 2005, a classification of crises is the first step to keeping them under control and allows for analysis and planning of crisis management actions. He defines four conditions for a good typology: 1) mutually exclusive classes, 2) exhaustive, covering also future events, 3) practicable, i.e., covering measures of prevention and 4) pragmatic, thus manageable.

Following Gundel: 2005, 110, we have typed crises as conventional, unexpected, intractable, and fundamental. In the case of conventional crises, the occurrence of the event is known and probable, thus predictable, and easy to prevent with proper quality controls and planning, e.g., the Peloponnesian War, especially as explained by Thucydides, was inevitable as Athens was on the rise and on a colliding course with the most iconic military power among Greek city-states, Sparta. Unexpected crises are less manageable. Once an unexpected and dangerous process has been triggered, it is almost impossible to stop it within a reasonable timeframe. An emergency response can combat the crisis successfully, but its surprising occurrence can hinder the solution.

[10] Originally from Greek krisis 'decision', krinō 'to decide' Liddel and Scott s.v. κρίνω, https://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0057%3Aentry%3Dkri%2Fnw

To illustrate this type of crisis, we cite the regime of the Thirty tyrants in classical Athens, the fragile pro-Spartan oligarchy installed in Athens after its defeat in the Peloponnesian War in 404 BCE. The third type is the intractable crisis that can have precedents in the past and be expected, but countermeasures are difficult because of the complexity of systems or conflicts of interest, e.g., the exile of Alcibiades, while being the leader of the Athenian fleet during the campaign in Sicily. Fundamental crises are unpredictable and difficult to influence because they give rise to chaotic, unprecedented circumstances. Examples of the fourth type are contagious illnesses, such as the plague in Athens in 430 BCE that led to a series of socio-political traumas.

6. Building the Ontoterminology of 'crisis'

The LACRIMALit methodology takes terms to be verbal designations of concepts in a given natural language, i.e., specific words that designate concepts, in compliance with the ISO principles on Terminology [IS0 1087; ISO 704]. This allows for extralinguistic modelling (conceptualisation) of crises independently of the different ways of talking about them in natural languages.

In computer science and information science, a computer-readable conceptualization of a domain is an ontology. There are different definitions of ontology [Guarino *et al*. 2009]. All of them rely on a formal knowledge model for the comprehensive description of a domain of knowledge that encompasses the set of concepts in the domain, their properties, and the relations that hold between concepts. Ontologies are used in practice for the representation of knowledge in a way that can be calculated by the computer, for the standardisation, semantic interoperability, knowledge discovery, complex question answering and automation of the inference process. In particular, the description of the properties of the objects of the world and their classification into categories (concepts), together with the description of the relations between these categories (concepts), enables further classifications of the objects, and the extraction of further associations between the concepts.

In order to represent the knowledge about the crises recorded in the corpus of ancient historians of the Graeco-Roman period in an interoperable manner, LACRIMALit will build a semantic resource combining an ontological component with a terminological one. Semantic interoperability between information systems is guaranteed, if and only if each can seamlessly carry out the tasks for which it was designed using data taken from the other. Ontologies are software artefacts whose purpose is to inject semantics into the data available on the Web, attention has turned toward the use of ontologies [Gruber 1993; Gruber 2009; Sowa 2000, Guarino *et al.* 2009; Staab & Studer 2009] for the representation of knowledge and for applications of automatic knowledge discovery. Through the incorporation of formal definitions, they also allow the application of basic inference mechanisms when interpreting data exploiting taxonomic and other relations built into the ontology.

There are several methods for building ontologies [Uschold and King 1995; Grüninger and Fox 1995; Fernández-López *et al.* 1997; Fernández-López 1999; Noy and McGuinness 2001; Corcho *et al.* 2003]. Some criteria set by these methodologies are: clarity, coherence, extensibility, etc.

The LACRIMALit project follows the ontoterminological approach which combines the semasiological and onomasiological approaches while taking into account the way of thinking of Humanists [Roche and Papadopoulou 2019; 2020]. An ontoterminology is a terminology (list of terms in natural language) whose conceptual system of the domain of interest is a formal ontology [Roche 2012]. In our ontoterminological approach, special attention is paid to the construction of the formal definitions of the concepts of the ontology.

6.1. Identifying terms

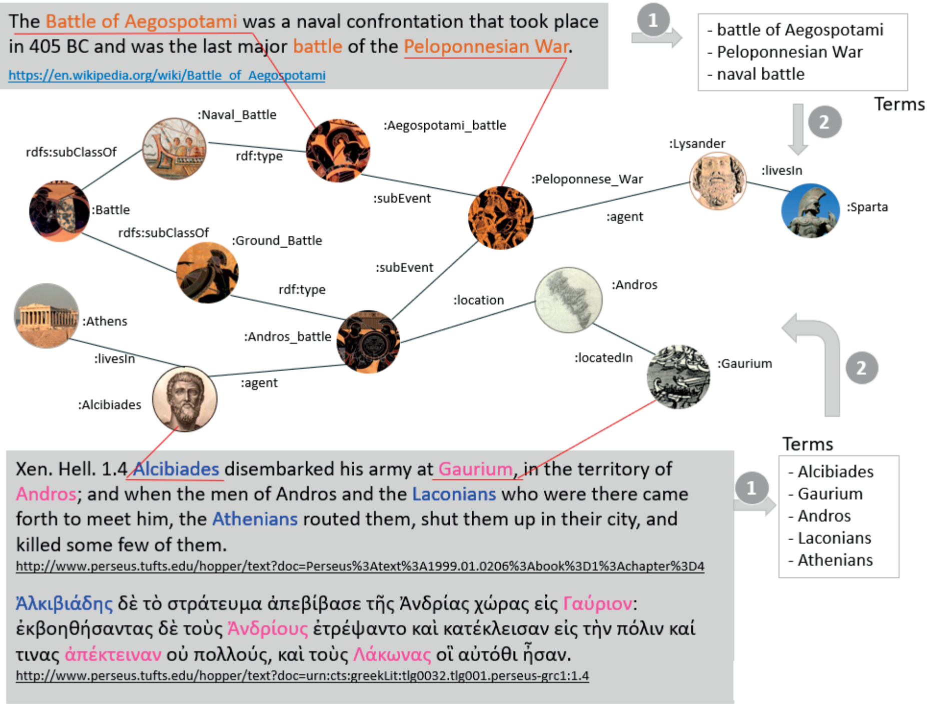
The terms were not extracted automatically from texts but provided by experts and illustrated by excerpts from the corpus. Terms are organised according to the denoted information: people, places, and events corresponding to as many corresponding concepts of the ontology.

The LACRIMALit project focuses on the following three basic categories of (political) crises are included in the typology and analysis:

a. emergency crisis incidents in times of war or peace, such as dispute (Gr. *διαφωνία, διαφωνέω*), military threat (Gr. *ἀπειλ-ή, -απειλέω*), etc., which usually require the undertaking of immediate measures
b. breach of trust between leaders and their followers; as well as the means (e.g., harangues, Gr. *Δημηγορία, λόγος*) by which leaders attempt to restore order
c. conspiracy (Gr. *συνωμοσία*), treason (Gr. *προδοσία*), revolt (Gr. *στάσις*), political confusion, tumult (Gr. *ταραχή*)

The example in Figure 2 shows how ontological data and information is drawn from the text of the Wikipedia article on the naval battle at

Aegospotamoi[11] and from the relevant passage from Xenophon's text from *Hellenica* book 1, paragraph 4[12]. The relevant terms denoting events such as are represented in the LACRIMALit model are expressed in the knowledge graph built from the information drawn from the texts.



The Battle of Aegospotami was a naval confrontation that took place in 405 BC and was the last major battle of the Peloponnesian War.
https://en.wikipedia.org/wiki/Battle_of_Aegospotami

Xen. Hell. 1.4 Alcibiades disembarked his army at Gaurium, in the territory of Andros; and when the men of Andros and the Laconians who were there came forth to meet him, the Athenians routed them, shut them up in their city, and killed some few of them.
http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0206%3Abook%3D1%3Achapter%3D4

Ἀλκιβιάδης δὲ τὸ στράτευμα ἀπεβίβασε τῆς Ἀνδρίας χώρας εἰς Γαύριον: ἐκβοηθήσαντας δὲ τοὺς Ἀνδρίους ἐτρέψαντο καὶ κατέκλεισαν εἰς τὴν πόλιν καί τινας ἀπέκτειναν οὐ πολλούς, καὶ τοὺς Λάκωνας οἳ αὐτόθι ἦσαν.
http://www.perseus.tufts.edu/hopper/text?doc=urn:cts:greekLit:tlg0032.tlg001.perseus-grc1:1.4

FIG. 2 – *Ontology-based annotation of text*

6.2. Building the ontology

There are different theories of concept [Roche 2015] defining as many approaches. The [ISO 1087] and ISO 704] principles on Terminology rely

[11] Available online: https://en.wikipedia.org/wiki/Battle_of_Aegospotami.

[12] Xen. *Hell.* 1.4 [Translation C.L. Brownson]. Available online via Perseus Digital Library https://www.perseus.tufts.edu/hopper/text?doc=Xen.+Hell.+1.4.&fromdoc=Perseus%3Atext%3A1999.01.0206.

on essential characteristics - a concept is defined as a unique combination of essential characteristics - whereas the main approach in knowledge engineering relies on the notion of class organising objects into sets according to their relationships. This article presents the first stage of the project which consists more in organising events into classes rather than defining terms. It is the reason why the second approach was chosen as well as the Protégé environment for building the LACRIMALit ontology. Protégé 3.3.1 [Musen 2015] is a free, open-source platform, a popular tool of Stanford University for developing Domain Ontology.

A LACRIMALit_Event is a type of action carried out by one or several LACRIMALit Agents that leads to changes of states in cultural, social or physical systems. It is made up of one or several LACRIMALit subevents, can have one or more causes and consequences as well as predecessors and successors, is located in a geographical space (Location), has a date of beginning and a date of ending. LACRIMALit crisis are subclasses of the LACRIMALit Event class.

The LACRIMALit Ontology is defined as a domain extension of some CIDOC classes. The LACRIMALit classes are organised into three main categories: Agent (including Group and Person), Event (Crisis) and Location, each of them defined as subclasses (rdfs:subClassOf) of respectively E39_Actor, E7_Activity, and E53_Place (see Figure 3). New relationships (object properties) between LACRIMALit_Events have been introduced for example to represent the causes and the consequences of an event.

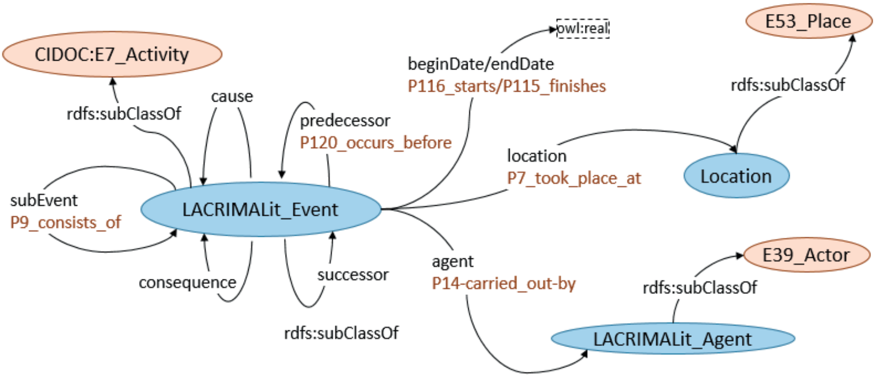
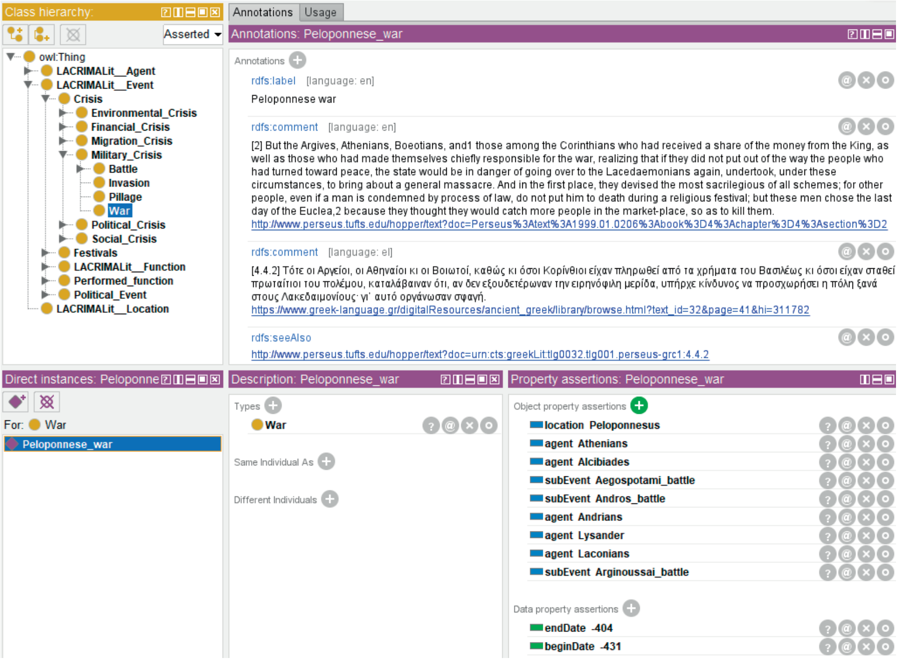


FIG. 3 – *LACRIMALit model as an CIDOC-CRM extension*

Figure 4 represents the formal description of the individual ‘Peloponnese war’ as an event whose begin date was 431 BCE and end date 404 BCE and that was composed of subevents (such as the battle of Aegos potami), that had locations such as Peloponnesus, and protagonists (groups such as the Athenians and the Laconians and persons such as Alcibiades and Lysander).



Annotations: Peloponnese_war

Annotations

rdfs:label [language: en]

Peloponnese war

rdfs:comment [language: en]

[2] But the Argives, Athenians, Boeotians, and1 those among the Corinthians who had received a share of the money from the King, as well as those who had made themselves chiefly responsible for the war, realizing that if they did not put out of the way the people who had turned toward peace, the state would be in danger of going over to the Lacedaemonians again, undertook, under these circumstances, to bring about a general massacre. And in the first place, they devised the most sacrilegious of all schemes; for other people, even if a man is condemned by process of law, do not put him to death during a religious festival; but these men chose the last day of the Euclea,2 because they thought they would catch more people in the market-place, so as to kill them.
http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0206%3Abook%3D4%3Achapter%3D4%3Asection%3D2

rdfs:comment [language: el]

[4.4.2] Τότε οι Αργείοι, οι Αθηναίοι κι οι Βοιωτοί, καθώς κι όσοι Κορίνθιοι είχαν πληρωθεί από τα χρήματα του Βασιλέως κι όσοι είχαν σταθεί πρωταίτιοι του πολέμου, καταλάβαιναν ότι, αν δεν εξουδετέρωναν την ειρηνόφιλη μερίδα, υπήρχε κίνδυνος να προσχωρήσει η πόλη ξανά στους Λακεδαιμονίους· γι’ αυτό οργάνωσαν σφαγή.
https://www.greek-language.gr/digitalResources/ancient_greek/library/browse.html?text_id=32&page=41&hi=311782

rdfs:seeAlso

http://www.perseus.tufts.edu/hopper/text?doc=urn:cts:greekLit:tlg0032.tlg001.perseus-grc1:4.4.2

Property assertions: Peloponnese_war

Object property assertions

location Peloponnesus
agent Athenians
agent Alcibiades
subEvent Aegospotami_battle
subEvent Andros_battle
agent Andrians
agent Lysander
agent Laconians
subEvent Arginoussai_battle

Data property assertions

endDate -404
beginDate -431

FIG. 4 – *Fragment of the LACRIMALit ontology in Protégé ontology editor*

6.3. Linking terms to concepts

The last stage consists in linking terms to the ontology. Since terms play a central role in semantic annotation to which a lot of information can be attached, they should be explicitly represented, i.e., as individuals of an OWL class ‘Term’ for example [Piccini 2015], and linked to individuals representing agents, places, events. However, terms corresponding to common nouns cannot be directly linked to classes since object properties are defined only between individuals. Classes can be treated as individuals, as it is allowed in OWL Full, using the same IRI to be both a Class (owl:Class) and an individual (owl:NamesdIndividual). Unfortunately, such an approach is not completely satisfactory: mixing knowledge of different types (domain, implementation) is difficult to understand and maintain. It is the reason why it was decided to go back to a simple representation of terms as labels in different languages attached to classes (rdfs:label, skos:prefLabel, skos:altLabel).

7. Evaluation

The last step is to evaluate the LACRIMALit ontology. Ontology evaluation is the task of measuring the quality of an ontology. Ontology evaluation is essential for wide adoption of ontologies in the Semantic Web and related technologies. There are different evaluation methods whose goal is "to assess the quality and correctness of the obtained ontology" [Sabou and Fernandez, 2012]. Criteria[13] allow to calculate the "richness" of an ontology such as the attribute richness[14] or relationship richness. Nevertheless, evaluation of criteria strongly depends on the aims of the ontology and the choices made for its implementation: "a good ontology does not perform equally well with regards to all criteria" [Vrandečić, 2009]. First of all, the ontology must allow providing the right answers to the competency questions. The competency questions have been translated into SPARQL[15] to query the OWL version of the LACRIMALit ontology built with Protégé. All of them are satisfied. Figure 5 presents the competency question "CQ4: Who are the protagonists of a war, in this particular case the Peloponnesian war?" translated in SPARQL and the results returned, i.e., the set of individuals to which the individual 'Peloponnese war' is linked by the 'agent' object property.

For the representation of facts such as "the Athenians took part in the Peloponnesian War" (i.e., facts whose Agent was a group), in Protégé, it has been required to use the same resource (IRI) both as an individual (Athenians as a protagonist) and as a class (Alcibiades is an Athenian).

[13] https://ontometrics.informatik.uni-rostock.de/wiki/index.php/Schema_Metrics.

[14] Attribute richness (AR) is defined as the average number of attributes (slots) per class. It is computed as the number attributes for all classes (att) divided by the number of classes.

[15] [SPARQL 2013] is a language dedicated to query knowledge graphs written in RDF-family languages.

SPARQL Query	protagonistName
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> PREFIX owl: <http://www.w3.org/2002/07/owl#> PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> PREFIX xsd: <http://www.w3.org/2001/XMLSchema#> PREFIX skos: <http://www.w3.org/2004/02/skos/core#> PREFIX foaf: <http://xmlns.com/foaf/0.1/> PREFIX lac: <http://o4dh.com/Ontologies/Crisis.owl#> SELECT?protagonistName WHERE { ?war rdf:type lac:War. ?war rdfs:label 'Peloponnese war'@en. ?war lac:agent?protagonist. ?protagonist foaf:name?protagonistName } ORDER BY?protagonistName	"Alcibiades"@en "Andrians"@en "Athenians"@en "Laconians"@en "Lysander"@en

Fig. 5 – *CQ 4 "Who are the protagonists of the Peloponnesian war" translated in SPARQL and the results it returned*

8. Conclusion

In this paper, we have presented LACRIMALit ontology, a model of concepts to organise historical knowledge about crises in the Graeco-Roman world and provide access to and understanding of these historical narratives. LACRIMALit is work-in-progress towards the semantic annotation that will enable the semantic querying of a vast number of ancient Greek texts. As such, it brings to the fore central common problems faced by digital humanists, especially those working with texts. For digital humanities work to fit into the framework of the semantic web and linked and open data, taking into account the way of thinking of domain experts, the following tasks are typically required: selecting a corpus of texts to study, defining the domain of knowledge one is interested in, create or choose an ontology for that knowledge domain, and formally annotate the relevant text passages using the ontology.

We have illustrated how the LACRIMALit ontology conceptualises crises in ancient Greek historiography and allows to answer the competency questions. We put particular emphasis on those essential terms that ancient historians use to present and discuss crises on the political scene, affecting the life of many and the course of subsequent events. We envision that populating the ontology with the crisis events from the whole corpus of ancient authors will provide a useful resource for digital historians: it can help historians to compare and contrast factual information about events.

The LACRIMALit ontology is a domain ontology defined as an extension of the CIDOC-CRM classes dedicated to the description of events involving one or more actors (E7 Activity). The LACRIMALit ontology was built using the Protégé environment, which allows the construction of ontologies in W3C format. If this environment is particularly well adapted to the organisation of individuals into classes, it is much less so with regard to the modelling of the linguistic dimension. The explicit representation of terms as individuals raises problems whose solutions are not really satisfactory.

In addition to the modelling issues necessitated by the theory underlying Protégé and the learning curve it presents for domain experts [Westerinen and Tauber 2017], the problem of knowledge and terminology modelling in digital humanities for the purposes of semantic annotation and knowledge retrieval remain open issues.

References

Angiolillo R., E. Elia, E. Nuti (eds) (2015). Crisi, Immagini, interpretazioni e reazioni nel mondo greco, latino e bizantino. Alessandria.

Arp R., B. Smith, A.D. Spear (2015). Building Ontologies with Basic Formal Ontology. The MIT Press.

Astrova I., A. Koschel, J. Lukanowski, J. L. Munoz Martinez, V. Procenko, M. Schaaf (2014) Ontologies for Complex Event Processing, World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:8, No:5.

Bekiari C. G. Bruseker, M. Doerr, C.-E. Ore, S. Stead, A. Velios (2021) Volume A: Definition of the CIDOC Conceptual Reference Model. Produced by the CIDOC CRM Special Interest Group, Version 7.1.1, April 2021, https://cidoc-crm.org/version/version-7.1.1

Björck, A. (2016). Crisis Typologies Revisited: An Interdisciplinary Approach. Central European Business Review. 5. 25-37. 10.18267/j.cebr.156.

Bringmann, K. (2003). Krise und Ende der römischen Republik (133-42 v.Chr.), Berlin.

Brown Windish S., C. Bonial, L. Obrst, M. Palmer (2021) The Rich Event Ontology: Ontological Hub for Event Representations. In T. Caselli,

E. Hovy, M. Palmer, & P. Vossen (Eds.), Computational Analysis of Storylines: Making Sense of Events (Studies in Natural Language Processing, pp. 47-66). Cambridge: Cambridge University Press. doi:10.1017/9781108854221.004

Casati, R. and A. Varzi (2020) [2002]. "Events", *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2020/entries/events/>.

Chiarcos C., P. Cimiano, J. Bosque-Gil, T. Declerck, C. Fäth, J. Gracia, M. Ionov, J. McCrae, E. Montiel-Ponsoda, M. P. di Buono, R.Saurí, F. Bobillo, M. Fazleh Elahi (2020). Pret-a-LLOD D5.1 Report on Vocabularies for Interoperable Language Resources and Services.

Coombs, T. (1998). An Analytic Framework for Crisis Situations: Better Responses from a Better Understanding of the Situation. Journal of Public Relations Research, 10(3), 177-191.

Corcho O., M. Fernández-López and A. Gomez-Perez (2003). "Methodologies, tools and languages for building ontologies. Where is their meeting point?", Data and Knowledge Engineering 46 (2003), no 1, p. 41-64.

Corda, I., B. Bennett, V. Dimitrova (2011). A Logical Model of an Event Ontology for Exploring Connections in Historical Domains. In M. van Erp, W. R.van Hage, L. Hollink, A. Jameson, R. Troncy, editors, *Proceedings of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011), Bonn, Germany, October 23, 2011.* Volume 779 of *CEUR Workshop Proceedings*, pp. 22-31.

Crane G. R. (ed.) (2012). Perseus Digital Library. Tufts University. http://www.perseus.tufts.edu

Cybulska A. K. and P. T. J. M. Vossen (2010) Event models for historical perspectives: Determining relations between high and low level events in text, based on the classification of time, location and participants. Published in Proceedings of 7th International Conference on Language Resources and Evaluation (LREC 2010), edited by Calzolari, N., Choukri, K., Maegaard, B., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., pp. 3355 - 3362

Doerr M. and A. Kritsotaki (2006). Documenting Events in Metadata. In M. Ioannides, D. Arnold, F. Niccolucci, K. Mania (eds) The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST http://www.cidoc-crm.org/sites/default/files/Documenting%20Events%20in%20Metadata.pdf

Engels, D. (2012). Le déclin: la crise de l'Union européenne et la chute de la République romaine: analogies historiques, Paris.

Fernández-López M. (1999). "Overview of Methodologies for Building Ontologies", in IJCAI-99 Workshop On Ontologies and Problem-Solving Methods (Stockholm, Sweden)

Fernández-López M., A. Gómez-Pérez and N. Juristo (1997). "METHONTOLOGY: from Ontological Art Towards Ontological Engineering''. In: Proceedings of the AAAI97 Spring Symposium, 1997, p. 33-40.

Franchet d'Espèrey, S. *et al*. (eds) (2003). Fondements et crises du pouvoir, Bordeaux. https://books.openedition.org/ausonius/7331?lang=en

Galton, A. (2012). The ontology of states, processes, and events. InterOntology 2012, 5: 35 - 45.

Getty Thesaurus of Geographic Names online (TGN) https://www.getty.edu/research/tools/vocabularies/tgn/

Gruber, T. R. (1993). "A Translation Approach to Portable Ontology Specifications." Knowledge Acquisition 5.2: 199-220.

Gruber, T. R. (2009). "Ontology." Encyclopedia of Database Systems. Springer-Verlag.

Grüninger M. and M. Fox (1995). "Methodology for the Design and Evaluation of Ontologies", Workshop On Basic Ontological Issues in Knowledge Sharing, Montreal, Canada.

Guarino, N., D. Oberle, and S. Staab (2009). "What is an Ontology?" In R. Studer and S. Staab (eds.) Handbook on Ontologies. Berlin Heidelberg: Springer. 1-17.

Guarino, N., R. Baratella, G. Guizzardi (2022). Events, their Names, and their Synchronic Structure

Gundel, S. (2005). "Towards a New Typology of Crises." Journal of Contingencies and Crisis Management, 13 (3), 106-115.

Heath, T. and C. Bizer (2011). Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology. Seattle, WA: Morgan & Claypool Publishers. doi:10.2200/S00334ED1V01Y201102WBE001

Herman, G. (ed.) (2011). Stability and Crisis in the Athenian Democracy, Stuttgart.

Ikuta, R., W. F. Styler IV, M. Hamang, T. O'Gorman, and M. Palmer. (2014). Challenges of adding causation to richer event descriptions. In Proceedings of Proceedings of the 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, pages 12–20, Baltimore, Maryland, USA, June 22-27, 2014. https://aclanthology.org/W14-2903.pdf

ISO 1087 (2019). Terminology work and terminology science-Vocabulary.

ISO 704 (2009). Terminology work - Principles and methods

ISO 21127 (2014). Information and documentation - A reference ontology for the interchange of cultural heritage information.

Jansen, L. (2019). Ontologies for the Digital Humanities: Learning from the Life Sciences? WODHSA. First International Workshop on Ontologies for Digital Humanities and their Social Analysis. http://ceur-ws.org/Vol-2518/paper-WODHSA5.pdf

Lagoze, C. and J. Hunter (2001). The ABC ontology and model. Journal of Digital Information. 2.

Liddell H. G. and R. Scott. 1940. A Greek-English Lexicon. revised and augmented throughout by. Sir Henry Stuart Jones. with the assistance of. Roderick McKenzie. Oxford. Clarendon Press.

Maienborn, C. (2011). Event semantics. In Semantics: An international handbook of natural language meaning. Vol. 33. Maienborn, Claudia, Klaus von Heusinger, and Paul Portner, eds., pp. 802-829. Walter de Gruyter,

McConnell, A. (2022). The Politics of Crisis Terminology. *Oxford Research Encyclopedia of Politics*.

Musen, M. A. (2015) The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003.

Noy N. F. and D. L. McGuinness, (2001). “Ontology Development 101: A Guide to Creating Your First Ontology”, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05, Stanford University, USA.

O’Gorman, T., K. Wright-Bettner, and M. Palmer. (2016). Richer Event Description: Integrating Event Co-reference with temporal, causal and bridging annotation. In Proceedings of 2nd Workshop on Computing News Storylines, pp. 47-56.

OWL Web Ontology Language Reference (2004). W3C Recommendation 10 February 2004, Document Status Update 2009. Editors: M. Dean and G. Schreiber, https://www.w3.org/TR/owl-ref/#Sublanguages

Piccini, S. (2015). “PLOTITERM: modélisation de la terminologie de Plotin en OWL.” TOTh 2015, Terminology & Ontology: Theories and applications. pp 313-343.

Raimond Y. and S. A. Abdallah (2006). The Event Ontology, http://motools.sourceforge.net/event/event.html

Ren, Y., A. Parvizi, C. Mellish, J. Z. Pan, K. Van Deemter, R. Stevens (2014). Towards competency question-driven ontology authoring. *European Semantic Web Conference*, pp. 752–767.

Roche, C. (2012). Ontoterminology: How to unify terminology and ontology into a single paradigm LREC 2012, Eighth international conference on Language Resources and Evaluation, Istanbul (Turkey), 21-27 May 2012, pp. 2626-2630.

Roche, C. (2015). “Ontological definition.” Handbook of Terminology, Volume 1, John Benjamins Publishing, 2015, pp.128-152.

Roche C. and M. Papadopoulou (2019). “Mind the Gap: Ontology Authoring for Humanists”. 1st International Workshop for Digital Humanities and their Social Analysis (WODHSA)- Episode V: The Styrian Autumn of Ontology, http://ceur-ws.org/Vol-2518/paper-WODHSA7.pdf

Roche C. and M. Papadopoulou (2020). “Rencontre entre une philologue et un terminologue au pays des ontologies”. Revue Ouverte d’Intelligence Artificielle, Volume 1, n°1, pp. 43-70

Rodrigues, F. H. and M. Abel. (2019). What to consider about events: A survey on the ontology of occurrents. Applied Ontology, p. 1-36. DOI: 10.3233/AO-190217

Rousseaux, F. and Lhoste, K. (2010). Towards a Collection-Based Knowledge Representation: The Example of Geopolitical Crisis Management. 10.5772/9538.

Sabou, M. and Fernandez, M. (2012). Ontology (network) evaluation. In Ontology engineering in a networked world (pp. 193–212). Springer.

Scherp, A., T. Franz, C. Saathoff, S. Staab (2009). F - A model of events based on the foundational ontology DOLCE+DnS ultralite. K-CAP’09 - Proceedings of the 5th International Conference on Knowledge Capture. 137-144. 10.1145/1597735.1597760.

Shaw, R., Troncy, R., & Hardman, L. (2009). LODE: Linking Open Descriptions of Events. *UC Berkeley: School of Information*. https://escholarship.org/uc/item/4pd6b5mh

Sowa, J. F. (2000). Ontology, metadata, and semiotics. Conceptual structures: Logical, linguistic, and computational issues: Springer, pp. 55–81. http://link.springer.com/chapter/10.1007/10722280_5

SPARQL (2013). SPARQL 1.1 Query Language W3C Recommendation 21 March 2013. Editors: S. Harris and A. Seaborne, https://www.w3.org/TR/sparql11-query/

Staab S. and R. Studer (2009). Handbook on Ontologies, second edition, Springer-Verlag, Berlin-Heidelberg.

Uschold M. and M. King (1995). “Towards a Methodology for Building Ontologies”, Workshop on Basic Ontological Issues in Knowledge

Sharing, held in conjunction with IJCAI-95, AIAI-TR-183, University of Edinburgh.

van Hage W. R., V. Malaisé, R. Segers, L. Hollink, G. Schreiber (2011). Design and use of the Simple Event Model (SEM), Journal of Web Semantics, Volume 9, Issue 2, pp. 128-136, ISSN 1570-8268, https://doi.org/10.1016/j.websem.2011.03.003.

Verba, S. (1965). 'Comparative political culture.' In L. W. Pye and S. Verba (eds.) Political Culture and Political Development. Princeton University Press, pp. 512-560.

Vrandečić, D. (2009). Ontology Evaluation. In: Staab S, Studer R, (eds.) Handbook on ontologies. Springer, pp. 293-313.

Westerinen A., R. Tauber. (2017) Ontology Development by Domain Experts (Without Using the "O" Word), Applied Ontology 12(2):1-13.

Summary (in French)

Cet article présente les travaux en cours menés dans le cadre du projet LACRIMALit pour la construction d'un modèle de crises pour l'historiographie du grec ancien et d'une terminologie des crises en grec (ancien et moderne) et en anglais. L'alignement des trois terminologies, grec ancien, grec moderne et anglais, repose sur une conceptualisation commune des crises. Celle-ci est représentée sous la forme d'une ontologie au format du W3C construite à l'aide de l'environnement Protégé et se présente comme une extension de classes CIDOC dédiées à la modélisation d'événements. Grâce à la modélisation des événements et de leurs relations tant temporelles que causales, il devient possible de lier les différentes ressources historiographiques facilitant ainsi leur parcours et le raisonnement pour répondre à des questions complexes.

Keywords

Ontology, Terminology, Modelling Events, Ancient Greek Historiography, Crisis

Evolution of Modular Ontology: Application to Personalization

Rahma Dandan*, Sylvie Despres*

* Université Sorbonne Paris Nord, LIMICS, F-93017, Bobigny, France
prenom.nom@univ-paris13.fr

Abstract. Ontology evolution is part of life cycle of ontology and represents a real challenge in the field of ontology engineering. Because of the variety of consequences caused by changes in an ontology (changes in requirements, modification of needs, evolution of modeled knowledge, etc.), special attention must be paid to the ontology evolution process in order to ensure its coherence and consistency. The strategy adopted depends on the extent of the modifications to be made to the designed ontology. The objective of this study is to present the steps leading to modular ontology evolution. We identify the knowledge that allows us to apply the necessary changes to the evolution of existing modules and the development of new ones. We present, through this evolution, concrete examples of inferences obtained to personalize the recommendation of food and activities according to the data collected on the elderly person.

1. Introduction

Ontology evolution is part of the life cycle activities of an ontology and represents a real challenge in the field of ontology engineering. Indeed, once the ontology has been designed, changes in requirements, modeled knowledge evolution, and correction of design flaws require the application of a number of changes that alter the ontology hierarchy. Thus, special attention must be considered to the ontology evolution process to ensure its consistency and coherence.

More and more studies support the interest and need to develop modular ontologies that are accessible and understandable by domain expert (Khan et Keet 2021). A modular ontology contains reusable, self-contained components that have defined relationships with other modular ontologies. Modularity represents an efficient solution to simplify the ontology development with a large and complex domain (d'Aquin *et al*. 2009). Moreover, the fact that each module has been designed to be self-sufficient maximizes both the reusability and the maintainability of the overall ontology (Jarrar 2005). Thus, the addition of new modules and the evolution of one or more modules can be done independently while generating the minimum of side effects.

The application context of this study is to prevent undernutrition and sedentary lifestyle. In France, approximately 2 million people suffer from undernutrition (HAS 2020) and more than 50% of the population over 65 years old is inactive (SPF 2020). Previously, the MIAM modular ontology was designed for suggesting "wellness" recipes for general population taking into consideration nutritional and sensory preferences of users (Despres 2016). To allow the elaboration of nutritional recommendations and activities according to data collected on the elderly (needs, pathologies, eating habits, sensory preferences, etc.), the MIAM evolution must be studied. This evolution is motivated by the changes of domain knowledge, applicative usage and collected data.

The objective of this study is to present the steps leading to the modular ontology evolution requiring the development of other context-specific modules. Before presenting the evolution process of MIAM, we present in the following section a synthesis of different approaches for ontology design and evolution.

2. Related Works

2.1. Definition of Ontology Evolution

There is no consensus on the definition of ontology evolution. Many definitions have been attributed to ontology evolution due to the complexity and diversity of the steps involved in this process. Among the selected definitions, ontology evolution is defined as:

> "the timely adaptation of an ontology to the arisen changes and the consistent management of these changes." (Haase et Stojanovic 2005);
>
> "may be caused by either a change in the domain, a change in the conceptualization or a change in the specification" (Flouris *et al*. 2006);
>
> "process (i) enables handling the required ontology changes; (ii) ensures the consistency of the underlying ontology and all dependent artefacts; (iii)

> supports the user to manage changes more easily; and (iv) offers advice to the user for continual ontology reengineering" (Stojanovic 2004).

These definitions show the complexity of this activity because of the variety of activities involved and the consequences of changes in requirements and the domain. Indeed, ontology evolution involves a dozen sub-domains: matching, alignment, articulation, translation, evolution, debugging, versioning, integration, merging, etc. (Wardhana, Ashari, et Sari 2018).

2.2. Ontology Evolution Approaches

Over the last 20 years, there has been a growing interest in the study of the ontology evolution process and a variety of approaches, techniques and tools have been proposed to ensure the success of the changes made. We present a synthesis of the different evolution approaches found in the literature. For this purpose, we are mainly interested in works presenting evolutionary steps.

(Stojanovic 2004) addressed the different steps of the evolution process and proposed an overview of the ontology evolution problem. Subsequently, (Flouris *et al*. 2008) provide a study clarifying these steps by focusing on the changes that was made to the ontology during its evolution. It also identifies research areas dealing with aspects of ontology change, and delineates their boundaries while providing terminological clarifications for each type of change. (De Leenheer et Mens 2008) explored approaches to ontology evolution designed in distributed and collaborative environments in which the dynamic aspect and the level of domain complexity is more important than single-user ontologies. More recently, (Kotis *et al*. 2020) has explored ontology engineering methodologies for the evolution and reuse of living ontologies especially in the context of collaborative ontology evolution. From his study, he emphasizes the importance of not neglecting the recommendations of ontology engineering methodologies for the evolution of living ontologies and provides recommendations based on the analysis performed. Each of these approaches brings a sometimes similar, sometimes different view of the ontology evolution activity. Although these approaches are motivated by the need to build an ontology evolution framework, they come from a variety of disjointed research domains making the understanding, adoption and application of evolution tasks more difficult.

To structure the synthesis of the different evolutionary frameworks found in the literature, we relied on the evolutionary cycle proposed by (Zablith *et al*. 2015) which distinguishes between these different terms. These authors synthesize the different stages found in the different studies within an evolutionary cycle following five stages, where each stage potentially relies on different inputs and contextual information represented in the circles: (1) The need detection stage is the starting point for any evolution. It detects if new concepts and relations should be added to the ontology, or if some ontology elements can be deleted; (2) The modification suggestion step allows to specify the concrete change operations necessary to make the ontology evolve; (3) The modification validation step filters out the changes that should not be added to the ontology. It ensures both that the ontology is logically coherent and consistent according to the specified constraints (formal validation) and that the domain changes are relevant; (4) The impact assessment step is performed according to the impact on external artifacts that depend on the ontology (ability to answer queries) or according to specific criteria (cost, benefits, etc.); (5) The change management step represents the step of applying, recording changes and tracking the different ontology versions.

To summarize the different evolution approaches found in the literature, we present in Figure 1 a comparative diagram of the different evolution steps for a panel of ontology evolution studies, based on the steps of the (Zablith *et al.* 2015) cycle. In this comparison, we distinguish between the phases of the evolution process and the change operations to be applied on the ontology (addition, deletion, modification) on the ontology entities (concept, property, instances and inheritance of concepts). This comparison highlights the inclusion of steps for each framework and the difference in vocabulary used. The majority of studies apply the process identified by (Stojanovic 2004). Very few studies show the consequences and impacts related to the application of change operations. (Safyan *et al.* 2019) identifies three possible change operations: adding specialized sub-concepts, adding extended sub-concepts, modifying existing concepts. Some studies identify changes to be made prior to evolution and consider the change management step as the beginning of ontology evolution (Haase et Stojanovic 2005). In the majority of evolution frameworks, the distinction between ontology management, modification, evolution, and versioning is, in some cases, unclear. The majority of evolution frameworks recognize the need for the ontology versioning task, however, they present this task as a separate activity, outside of the evolution task. (Flouris *et al.* 2008) consider this task outside of the ontology evolution process on one hand and as a task for change propagation on the other. The advantage of the evolution cycle proposed by (Zablith *et al.* 2015) is that it includes the versioning task in the ontology evolution tasks. Moreover, it offers a higher granularity to evolve ontologies. Indeed, for each change made on the ontology, the application of the cycle steps is performed in an iterative way. Finally, their cycle includes the functionalities associated to each step (circles), often considered outside the cycle.

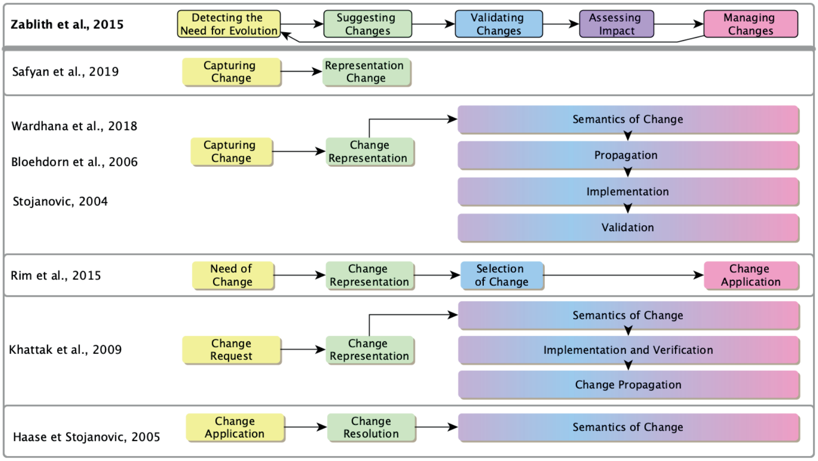


FIG. 1 – *Comparison of the evolution stages in comparison to the cycle stages proposed by (Zablith et al. 2015).*

The synthesis allows us to position ourselves within the existing approaches and to identify which approach best meets our evolution needs. In general, the strategy adopted to evolve an ontology depends on the extent of the modifications to be made to the ontology to be evolved. In order for the chosen approach to support ontology evolution, it is necessary to clearly identify the tasks and subtasks that underlie this evolution. For this, in addition to the evolution steps, the possible change operations according to the ontology entities are interesting to take into consideration when they are specified. In view of the evolution cycle granularity proposed by (Zablith *et al.* 2015) and the inclusion of the version management process, we have relied on this cycle steps to evolve MIAM.

3. Evolution of MIAM Ontology

3.1. Step 1: Detecting the Need for Evolution

The needs detection step represents the first step of ontology evolution and requires knowledge acquisition. The sources used are based on unstructured knowledge (raw text, diagrams, tables, etc.) mainly from official sources (PNNS, HAS, WHO, etc.) or specialized health structures (HUG, etc.). MIAM evolution is motivated by the following changes.

3.1.1. Changing PNNS Nutritional Recommendations

MIAM was designed based on the PNNS 3 (French National Nutrition and Health Program) recommendations and on the nutrition expertise of the Nutritional Epidemiology Research Unit (UREN). Since then, the recommendations have evolved from PNNS 3 to PNNS 4. The goal of the PNNS evolution was to simplify messages, to be more interactive and to include neglected foods. This update led us to modify some existing concepts in MIAM. In new version of PNNS, three additional food families have been created (pulses, cold cuts, nuts), some existing families have been modified (wholemeal starch, fish), and some families "to be limited" have been grouped together (fatty, sweet, salty and ultra-processed products). In some cases, like this one, it becomes more obvious to create again the concept describing new PNNS families rather than applying a series of change operations on existing concepts.

The vocabulary used to designate the food categories is not always identical according to the sources. Thus, particular attention has been paid to this point to avoid introducing semantic inconsistencies. A semantic inconsistency occurs when the meaning of an entity changes during the evolutionary process (Tamma et Bench-Capon 2001). In addition to vocabulary changes used, recommendations have also evolved. For example, the recommendation for dairy products has changed from 2 per day for adults to 3 or 4 per day for the elderly. These differences in recommendations require reworking a number of concepts describing the PNNS food groups and recommended portions to define them according to the individual's profile.

3.1.2. Changing Data Collection Needs

Data collection needs are conditioned by the specific needs of the elderly. To meet their needs, data must be collected on their health status, preferences (dietary, sensory, etc.), constraints, context, environment (social, family, etc.), level of dependence, objectives, motivations, barriers, etc. The collection tools used to build the person's profile are smart objects and questionnaires (Figure 2). The data collected can be of different types: raw or calculated. Each type of data requires a different use, collection and therefore processing. Among these treatments, calculations can be made from certain data. Thus, from the raw data collected, a certain number of data can be exploited to complete the profile. The data collected from the elderly are of different types: when they come from sensors, they are measured; when they come from questionnaires, they are declared; and when they are analyzed, whatever their source, they are calculated. From these means of collection, the objective is to be able to exploit this data to deduce new knowledge and personalize the recommendation. From the smart objects, data related to the activity profile (number of steps, time of inactivity, etc.), to the health profile (blood sugar, etc.) are collected. From the online questionnaires, all data considered useful to personalize the recommendations can be taken into consideration and are related to the cases treated in this study. Moreover, good recommendations can only be obtained when an accurate model of the users and their needs is available. To meet the nutritional needs of the elderly, they must first be identified. This requires a definition of person that is as complete as possible. The characterization of person according to the different aspects meeting the needs of application, consists in defining all the data used to build the profile (Figure 2).

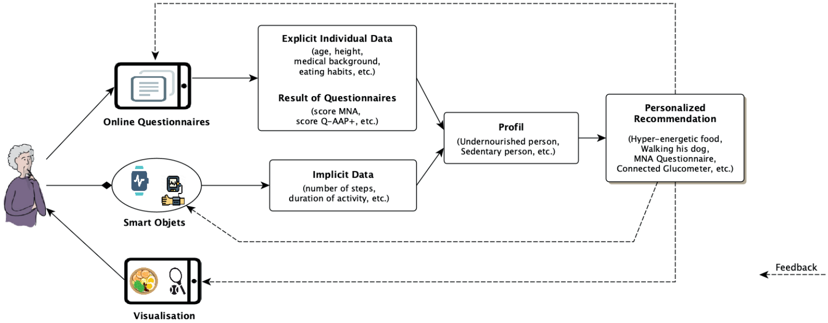


FIG. 2 – *Process of data exploitation.*

3.1.3. Change in Application Usage and Recommendations

Identifying changes requires knowledge acquisition. This knowledge concerns: the specific needs of the elderly; the data and elements on which it is possible to act in order to target the recommendation ontology entities the elements allowing the construction of profile; the exploitable and specific nutritional recommendations for a studied pathology, a sign, a symptom, etc. For example, to delay the onset of age-related health concerns (osteoporosis, hypertension, etc.) and the onset of nutritional disorders related to health concerns (undernutrition, chewing disorders, etc.), nutritional needs must be identified in order to be met. However, it is not easy to get people to adopt nutritional recommendations and change their behavior and eating habits so that their diet can cover their energy and nutritional (protein, vitamins, etc.) needs. Thus, the recommendation of foods rich in energy or protein will meet the nutritional needs of the malnourished elderly.

In order to age better, the health factors involved in the modification of diet or activity must be considered for a better management of health status of the elderly. Some of pathologies (diabetes, cancer, treatments, etc.) can disrupt nutritional status and thus cause undernutrition, weight loss or, on the contrary, excess weight. The pathologies require a diet modification have therefore been taken into consideration in our design to personalize the recommendations. The pathologies included are mainly those for which a recommendation is available. For example, people with diabetes and hypertension should limit the consumption of high salt products.

Undernutrition is an insidious and inconspicuous process. To fight against it, early detection of undernutrition signs is essential. For this, the detection of the causes (low BMI, loss of appetite, etc.) helps to orient the recommendation. The different management strategies consist of: enriching the diet, adapting the diet to digestive disorders, and improving the context of meal intake. We also looked at dietary enrichment. It consists in enriching the traditional diet by increasing the protein and calorie content of the components (starters, main courses, dairy products, desserts, snacks, drinks, etc.) of the meal without increasing their volume. Different basic products can have this effect (milk powder, whole condensed milk, etc.). All these foods are available in MIAM and can be exploited to specify diets.

The nutritional recommendations that we seek to provide concern recommendations associated with a food, a food group of the PNNS, a diet. Thus, from the different sources exploited, the recommendations retained take into consideration the health problems encountered by the elderly. The recommendation suggested to the user can be presented in the form of a list: of activity or activity program: to encourage him/her to practice physical activities related to his/her health status, preferences, etc.; food or nutritional diets: to enrich their diet; personalized questionnaires from short self-assessment tests: to target health determinants and build the gerontological profile of the elderly. Finally, in order to provide a personalized recommendation adapted to the individual needs of each person, the data of the elderly concerning the different terms found in the assessment questionnaires (health status, nutritional and activity status, etc.) are exploited to complete the data associated with the person. For example, in the case of undernutrition, we initially seek to detect the signs of undernutrition in order to prevent its occurrence and thus personalize the recommendations. Therefore, we have included in our application using the behavior of the application in relation to the collected data. When the person indicates a weight and a height allowing a BMI lower than normal, then the Mini Nutritional Assessment questionnaire (MNA), can be proposed to the elderly (Kaiser *et al.* 2009). The inclusion of system behavior in our study allows us to improve the data collection stage through the construction of personalized online questionnaires.

In summary, the knowledge acquired for the identification of changes concerns: the specific needs of the elderly; the data and elements on which it is possible to act to target the recommendation; the elements allowing the construction of profile; the exploitable and specific nutritional recommendations for a studied pathology, a sign, a symptom, etc.

3.2. Step 2: Suggesting Changes

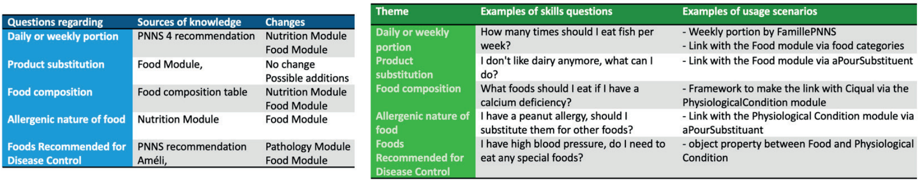
The suggestion of changes allows us to identify the changes to be made on the modules concerned by the evolution.

3.2.1. Competency Questions and Scenario

From the acquisition of needs, competency questions and scenarios (Grüninger et Fox 1995) have been defined. Traditionally, competency questions are used in ontology development to gather functional requirements of users in specific use cases ensuring that all relevant information is encoded (Jacobson *et al.* 1992). Competency questions are thus a means to determine the specifications of ontology to be designed and to evaluate a knowledge model. They consist of a set of questions that an ontology should be able to answer according to scenario.

The competency questions that have been established make it possible to clarify the needs previously identified in order to make the ontology evolve and correspond to questions expressed by the elderly or the caregivers taking care of them. They are constructed in such a way as to answer the elderly's questions concerning the foods that are "to be favored" to consume or, on the contrary, "to be limited" according to the elderly person's needs. These needs may be related to sensory preferences or to a diet to be followed. These different aspects must therefore be taken into consideration in order to arrive to relevant recommendations. Let's take as an example the following question: "How many times a day did you consume dairy products (milk, sugar-free yogurt or low-fat cheese)? If the answer is less than 2 times a day, we must be able to include all the products in the DairyProduct class into the recommendations so that the person can make a choice according to their own preferences.

Competency questions leading to MIAM evolution were categorized according to the themes addressed and used to define the modules involved in the changes (Figure 3). This filtering shows an overview of the competency issues and changes to be made on the different MIAM modules involved in the evolution. From the skill issues identified from forums or frequent problems, scenarios were built to determine the specifications of the changes to be applied to MIAM (Figure 3). These scenarios describe what the ontology should compute to answer the competency questions. During the exploration phase of the different resources available to answer the competency questions, we can quickly notice the variety of vocabulary used to designate the same food groups. To remove this ambiguity leading to inconsistencies, all terms were listed so that for each preferred term (in this case the term used by the PNNS), the corresponding alt-labels could be defined. For example, the concept "Fat" was replaced by the concept "Rapeseed, walnut, olive oil" named as such in the PNNS.



Questions regarding	Sources of knowledge	Changes
Daily or weekly portion	PNNS 4 recommendation	Nutrition Module Food Module
Product substitution	Food Module,	No change Possible additions
Food composition	Food composition table	Nutrition Module Food Module
Allergenic nature of food	Nutrition Module	Food Module
Foods Recommended for Disease Control	PNNS recommendation Améli,	Pathology Module Food Module

(a)

Theme	Examples of skills questions	Examples of usage scenarios
Daily or weekly portion	How many times should I eat fish per week?	- Weekly portion by FamillePNNS - Link with the Food module via food categories
Product substitution	I don't like dairy anymore, what can I do?	- Link with the Food module via aPourSubstituent
Food composition	What foods should I eat if I have a calcium deficiency?	- Framework to make the link with Ciqual via the PhysiologicalCondition module
Allergenic nature of food	I have a peanut allergy, should I substitute them for other foods?	- Link with the Physiological Condition module via aPourSubstituant
Foods Recommended for Disease Control	I have high blood pressure, do I need to eat any special foods?	- object property between Food and Physiological Condition

(b)

FIG. 3 – *Overview of competency questions filtering leading to MIAM evolution (a) and examples of competency questions and scenarios according to the theme (b).*

This step allowed us to identify the semantic relations as well as the concepts on which we must apply the changes. For example, we will define in the Person module the knowledge needed to collect data (physiological, nutritional needs, dietary habits, etc.) and recommend foods (or diets), activities (or activity programs) or questionnaires according to the specific needs of the elderly. Modelling modifications of some of the MIAM concepts were essential in order to consider, in particular, the change in the vocabulary used following the evolution of nutritional recommendations provided by the PNNS. The identification of the changes allowed us to define the MIAM modules concerned by the evolution.

3.2.2. Presentation of MIAM

MIAM is a modular ontology composed of a core module, the Person module and the Food module, which is used by the 7 other thematic modules (Figure 4): Nutrition (specific to nutrition); Cooking (relating to the realization of recipes and links with the dish types); Preparation (relating to the basic preparations associated with a recipe); Unit (relating to the cooking domain metrics (coffee spoon, mustard glass) and international metrics (gram, milliliter, etc.); Material (relating to materials used to realize the recipes) and Sensorial (relating to the sensory aspects characterizing the food and recipes). The basic concepts linking the general concepts of MIAM are gathered in “CORE-ONTO”. Many modules are interesting to reuse, like the Sensory module, and will be the subject of a future study.

3.2.3. Modules Targeted by Evolution

In order to meet the needs for physical activity recommendations and data collection from questionnaires and smart objects, MIAM evolution required: the evolution of certain MIAM modules and the development of new modules. In this study, the MIAM modules concerned by the evolution are: Food, Nutrition and Person (Figure 4). The new modules designed are the modules: Activity, Pathology, Sign, Symptom, SmartObject and Measure. All the imported modules (dark blue arrows) form ONAFE (Ontology of Nutrition and Activity For Eldely). To suggest changes, we did not limit ourselves to the content available in the analyzed documents. We have extended our reflection to consider the knowledge previously modelled in MIAM. We expose the suggestions of changes by defining competency questions and scenarios of use.

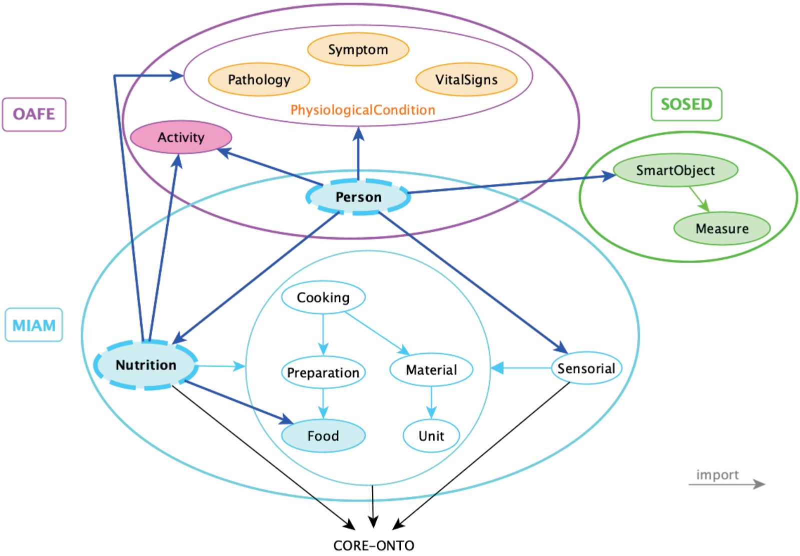


FIG. 4 – *Design showing the modules of MIAM involved in the evolution and the new modules to be developed to design ONAFE. The modules that have undergone an evolution are in dotted line.*

3.3. Step 3: Validating Changes

The first step in suggesting changes begins with the extraction of modules affected by the evolution.

3.3.1. Module Importation

The import of the modules is done in Protégé. The Person module is a module to which the other modules refer. The direction of the arrows indicates which module imports another module. Figure 4 shows that the person is at the center of our design. Indeed, the Person module imports all the other modules (except Food). Although the Food module is concerned by the evolution, it is imported only by Food. Each module contains concepts, relations and constraints (included in the T-Box) related to each domain and are then composed to form a unique ontology (unique T-box). The orientation of the arrows indicates the import direction. For example, the Nutrition module must bring the Activity and Food modules. The relations linking to the imported modules must be defined in the main module.

The modules concerned by the evolution were initially extracted from MIAM. From the modules selected in MIAM and requiring an evolution, we defined the changes to be made on the modules and the links between the modules. The new conceptualization takes into consideration the evolution of changes.

3.3.2. Evolution of Modules Food and Nutrition

The Food module is characterized as a stand-alone module and has not undergone any major changes. It was mainly used to characterize foods according to their nutritional value (HyperenergeticFood, HighProteinFood, etc.) and to define which foods are “to be favored”, “to be limited” and “to be avoided” in order to include them in a specific diet (UndernourishedDiet, DiabeticDiet, etc.) according to the PNNS groups. Thus, according to each nutritional need, a diet will be suggested.

The diets included in the Nutrition module currently concern diets related to identified health problems (undernutrition, diabetes, hypertension, etc.) and for which a recommendation can be personalized. The evolution of the Nutrition module cannot be done independently of the Person module. In the case of undernutrition, high protein and hyperenergetic foods can be recommended. In the case of a diabetic person, foods with a low glycemic index can be recommended. The Nutrition module contains all the concepts related to nutrition such as food allergens, meal structure (FoodAllergen, Meal, etc.). These concepts will then have to be integrated into the diets.

The Nutrition module also contains other MIAM modules not required for application use. Decisions on whether to remove a concept not included in the new requirements are made on a case-by-case basis. For example, for GroupePNNS3 the decision depends on the evolution of needs. It may be used, for example, to determine a food group used by other recommendations based on the PNNS 3 food groups.

3.3.3. Evolution of Module Person

The Person module is a module to which the other modules refer. It originated in MIAM and has evolved through the evolutionary stages to include knowledge of the new requirements for collecting, recommending, and profiling the person. To perform this evolution task, the representation of the elderly person, based on the data collected about him/her, provides an overview of the terms related to the Person and the variety of modules involved in its evolution (Figure 5).

The Person module was completed to characterize the person based on the data collected and evolved as the competency questions were answered to address the changes previously defined. The concepts represented in Person are used to build a profile of the older person and to personalize recommendations. The data collected is either fixed or variable. A person's fixed data (Name, Height, etc.) is acquired only once and does not vary over time. Some properties can either vary over time (Weight, etc.) or be contextual (Heart Rate, Blood Pressure, etc.). Thus, all the variables involved in the definition of the health state (Sign, Symptom, Pathology), the activity profile and the social environment (SocialCharacteristic, etc.), the goals that the person wishes or should reach (Goal) have been described. This modelization was enriched by associating the barriers related to the health status, the activity or social profile (Alfaifi *et al*. 2017) of the elderly person.

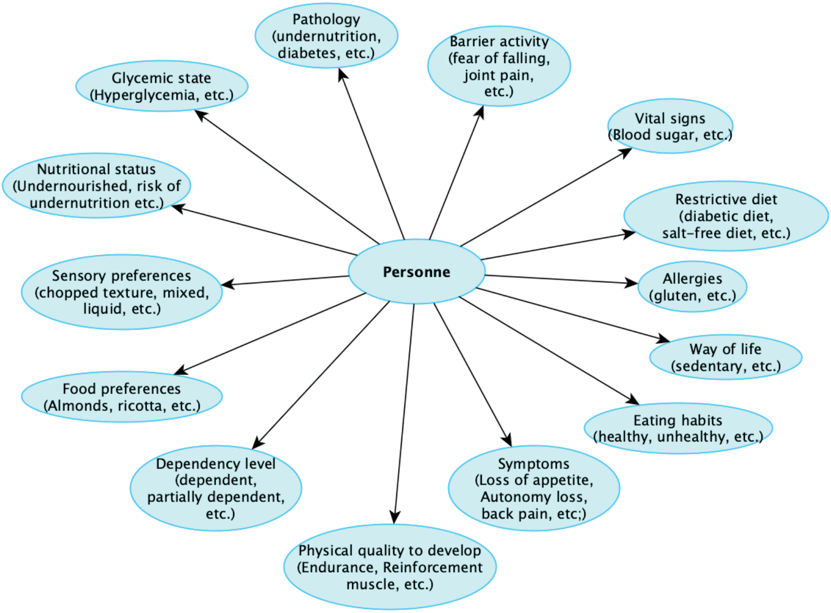

Personne
Pathology (undernutrition, diabetes, etc.)
Barrier activity (fear of falling, joint pain, etc.)
Glycemic state (Hyperglycemia, etc.)
Vital signs (Blood sugar, etc.)
Nutritional status (Undernourished, risk of undernutrition etc.)
Restrictive diet (diabetic diet, salt-free diet, etc.)
Sensory preferences (chopped texture, mixed, liquid, etc.)
Allergies (gluten, etc.)
Way of life (sedentary, etc.)
Food preferences (Almonds, ricotta, etc.)
Eating habits (healthy, unhealthy, etc.)
Dependency level (dependent, partially dependent, etc.)
Symptoms (Loss of appetite, Autonomy loss, back pain, etc;)
Physical quality to develop (Endurance, Reinforcement muscle, etc.)

FIG. 5 – *Conceptualization of the Person module to meet evolutionary needs*

3.4. New Modules Engineering

The study presented in this section is based on a preliminary study carried out for ontology design of OAFE (Ontology Activity For Elderly) (Dandan *et al*. 2018) was designed to model the activity domain. In this study, we present the knowledge acquisition stage and list a non-exhaustive list of skill questions and scenarios mainly from use cases identified from forum discussions.

NeOn (Suárez-Figueroa *et al*. 2012) constitutes the methodological framework in which we situate ourselves to build the different modules. Scenarios 1 (Specification, Scheduling, Conceptualization, Formalization, Implementation) and 7 (Ontology Design Patern Reuse) have been applied. The scope of the ontology is established by adopting the methodology of (Uschold et Gruninger 1996).

3.4.1. Ontology Reuse

When developing a new ontology, it is recommended to reuse existing ontologies as much as possible. When appropriate ontologies exist, the new ontology should start by importing higher level ontologies. This simplifies the development since one can focus only on domain or application specific knowledge. Activity is an important concept in many domains (geography, transportation, psychology, etc.), and a number of activity-related ontologies have been developed. The W3C Provenance Working Group (PROV), meanwhile, provides a minimal PROV-O ontology (PROVenance Ontology, 2013) consisting of three core concepts (Activity, Entity and Agent) and is an interesting resource to reuse. An Activity ODP also exists and is more comprehensive (Abdalla *et al*. 2014). It provides a generic ontology design pattern to model the common core of the Activity module in different domains. Finally, a number of ontologies have been designed related to pathologies, signs and symptoms (geriatrics, Human Disease Ontology, etc.).

The decision on whether to reuse an ontology or a vocabulary depends on several factors. The ontologies cited were not reused either because their context was too distant or because of the difficulty of reusing a vocabulary that required translation. For example, the Human Disease Ontology could have been reused to define the main classes of pathologies, however, ontology is voluminous and requires considerable work. It would be interesting to exploit tools to automate the translation of all the ontology entities. During the reuse scenario, only the Activity ODP was chosen to be reused in the design of Activity module. The representation of the ODP was reused and the different reused concepts and relations (Activity) were translated. The other ontologies cited were not reused either because their context was too distant, due to the lack of model granularity or the absence of the ontological resource.

3.4.2. Activity and PhysiologicalCondition Modules

The Activity module was designed for describes the activity characteristics and the context in which it takes place. The conceptual model showing the first and second level relationships related to the Activity concept has presented in (Dandan *et al*. 2018).

The PhysiologicalCondition module gathers three classes: Pathology, VitalSigns, Symptom. It contains the main pathologies and symptoms encountered in the elderly. The vital signs are defined according to threshold values allowing to characterize them as low, normal, high, etc. Among the signs, we distinguish the vital signs because of their measurability by the smart objects currently available on the market. Vital signs are the four signs that can give an immediate measure of the overall functioning and health status of the body (temperature, heart rate, etc.). The ranges of these measurements vary according to age, weight, gender and general health.

3.4.3. SmartObject and Measure Modules

The SmartObject and Measure modules constitute the ontology of SOSED (Objects for collected hEalth SEnsor Data). It has been designed to describe the smart objects that can be used by the elderly person to monitor health and activity parameters. The SmartObject module includes the smart objects that could accompany them in their daily life and meet the need for data collection and identification of signs of undernutrition and sedentary lifestyle. Thus, smart objects can be proposed according to the established profile. For example, in the case of a diabetic person, a connected glucometer will be recommended to facilitate the monitoring of his or her blood sugar level and to personalize the nutritional recommendations according to it. In general, foods with a high glycemic index are “to be limited” and foods with a low glycemic index are “to be favored”. When the person’s blood sugar is very high, foods with a high glycemic index should “to be avoided” and foods with a medium glycemic index should “to be limited”. The Measure module includes measurements from smart objects identified as useful for the application. For example, the connected watch can measure the number of steps. In the case of a sedentary person, when the average number of steps per day is below the recommendations, low intensity activities will be suggested. For the moment, SOSED ontology complies with the design perimeter. However, it could evolve later to include other modules according to the evolution of the application needs.

3.5. Step 4: Assessing Impact

Like MIAM, the ONAFE modules have been designed in a modular way following a compositional approach. In Protégé, the import relation was used to build ONAFE which represents all the modules involved in the evolution tasks presented. The modelization of the different conceptual maps was used as a basis for the evolution and construction of the modules formalized in OWL2. The concepts of the conceptual model are represented by classes and the relations by properties (object property, data property). Each entity is annotated by a label, an altlabel, and a definition. The ONAFE evaluation has been iterative. The application of the previously established scenarios made it possible to simplify the ontology management during its evolution, and to control the impact of its evolution, notably by establishing a series of DL Queries. As new modules evolved and were designed, individuals were introduced to visualize the classifications obtained and the inferences expected by the reasoner. Changes were managed in a way that ensured progression and traceability between different versions of the ontology. In the end, all changes made to an ontology were integrated in order to maintain its structure, consistency and coherence.

3.6. Step 5: Managing Changes

The version management was carried out in an iterative way. For each change applied to a module, an operational version is generated with the annotations corresponding to the changes made and then renamed according to modification date. Collaborative work requires taking into consideration the important aspects related to sharing and updating documents and modules related to the application of each step of the evolution cycle. The use of online collaborative workspaces (Git, Overleaf, etc.) facilitates the application of this version management step.

4. Conclusion

Through this study we presented the steps leading to modular ontology evolution, describing the knowledge allowing to suggest food and personalized diets according to the nutritional needs of elderly. We demonstrate through this evolution case, the interest of designing modular ontologies to manage the evolution and development of ontologies. The cycle on which we rely has many advantages. In addition to detailing the evolution steps, the rigor in the application of this cycle is important and allows us to manage the different changes in a successive manner.

Although we have based ourselves on a global approach of ontology evolution, MIAM evolution presents a certain number of aspects linked to the ontology modular architecture to be evolved. Each module corresponds to a defined domain space. The advantage offered by the modularity of ontologies is that it favors the reuse and the ontology maintenance throughout its evolution and the integration of other ontologies at the time of MIAM evolution. Each modular ontology represents domain concepts that can be adapted to specific needs, person-centered design and evolution, and recommendations to be provided. The specific user requirements are taken into consideration in the definition of competence issues and usage scenarios. This methodology supports the evolution of modules and the evaluation of the final ontology. The application of the evolution steps in an iterative manner, according to the needs identified throughout the process, makes it possible to resolve the design problems encountered and to validate the version generated at the end of each task.

More and more studies support the interest and the need to design modular ontologies, accessible and understandable by domain experts. The fact that each module is designed to be self-sufficient has facilitated the evolution of modules and the design of new modules, maximizing both the reusability and the maintainability of the overall ontology. Thus, the addition of new modules and the evolution of one or more modules could be done independently, generating the minimum of side effects and not affecting the concept hierarchy of the other modules. The main motivation for modular ontologies is scalability, complexity management, comprehensibility, context awareness, customization and reuse. A sufficiently modularized ontology is designed so that it can easily be adapted to each use case. In the ONAFE ontology, the majority of the modules are independent and the relationships dependent on other modules are defined in the module that matters the targeted module. In this way, evolution is facilitated and dependencies between modules are only broken when the imported module is no longer essential to the application. It can nevertheless be used for other applications.

Reference

Abdalla, Amin, Yingjie Hu, David Carral, Naicong Li, et Krzysztof Janowicz. 2014. « An ontology design pattern for activity reasoning ». 78-81. CEUR-WS.org.

Alfaifi, Yousef, Floriana Grasso, et Valentina Tamma. 2017. « Towards an Ontology to Identify Barriers to Physical Activityfor Type 2 Diabetes ». In *Proceedings of the 2017 International Conference on Digital Health*, 16-20. DH '17. NY, USA: ACM.

Aquin, Mathieu d', Anne Schlicht, Heiner Stuckenschmidt, et Marta Sabou. 2009. « Criteria and Evaluation for Ontology Modularization Techniques ». In *Modular Ontologies: Concepts, Theories and Techniques for Knowledge Modularization*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer.

Dandan, R., S. Despres, et J. Nobécourt. 2018. « OAFE: An Ontology for the Description of Elderly Activities ». In *2018 14th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, 396-403.

De Leenheer, Pieter, et Tom Mens. 2008. « Ontology Evolution ». 131-76.

Despres, Sylvie. 2016. « Construction d’une ontologie modulaire. Application au domaine de la cuisine numérique », Revue d’Intelligence Artificielle, 30 (5): 509-32.

Flouris, Giorgos, Dimitris Mankanatas, Haridimos Kondylakis, Dimitris Plexousakis, et Grigoris Antaniou. 2008. « Ontology Change: Classification and Survey ». *The Knowledge Engineering Review* 23 (juin): 117-52.

Flouris, Giorgos, Dimitris Plexousakis, et Grigoris Antoniou. 2006. « Evolving Ontology Evolution ». In *SOFSEM 2006: Theory and Practice of Computer Science* 3831:14-29. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin, 1995. « Methodology for the Design and Evaluation of Ontologies ».

Haase, Peter, et Ljiljana Stojanovic. 2005. *Consistent Evolution of OWL Ontologies. Lecture Notes in Computer Science*. Vol. 3532.

Haute Autorité de Santé (HAS), Laëtitia. 2020. « Diagnostic de la dénutrition de la personne âgée ». RECOMMANDER LES BONNES PRATIQUES.

Jacobson, Ivar, Magnus Christerson, ACM Press Staff, Patrik Jonsson, et Gunnar Övergaard. 1992. *Object-Oriented Software Engineering: A Use Case Driven Approach*.

Jarrar, Mustafa. 2005. « Towards methodological principles for ontology engineering ».

Kaiser, M. J., J. M. Bauer, C. Ramsch, W. Uter, Y. Guigoz, T. Cederholm, D. R. Thomas, *et al*. 2009. « Validation of the Mini Nutritional Assessment Short-Form (MNA-SF): A Practical Tool for Identification of Nutritional Status ». *The Journal of Nutrition, Health & Aging* 13 (9): 782-88.

Khan, Zubeida, et C. Keet. 2021. « Structuring Abstraction to Achieve Ontology Modularisation ». In book: *Advanced Concepts, Methods, and Applications in Semantic Computing*, 72-92.

Kotis, Konstantinos I., George A. Vouros, et Dimitris Spiliotopoulos. 2020. « Ontology Engineering Methodologies for the Evolution of Living and Reused Ontologies: Status, Trends, Findings and Recommendations ». *The Knowledge Engineering Review.*

Lockwood, Craig, Tiffany Conroy-Hiller, et Tamara Page. 2004. « Vital Signs ». *JBI Reports* 2 (6): 207-30.

Safyan, Muhammad, Zia Qayyum, Sohail Sarwar, Muddesar Iqbal, Raúl García Castro, et Anwer Al-Dulaimi. 2019. « Ontology Evolution for

Personalized and Adaptive Activity Recognition». *IET Wireless Sensor Systems* 9 (mars).

Santé Publique France (SPF). 2019. «Activité physique et sédentarité dans la population française. Situation en 2014-2016 et évolution depuis 2006-2007».

Stojanovic, Ljiljana. 2004. «Methods and tools for ontology evolution».

Suárez-Figueroa, Mari Carmen, Asunción Gómez-Pérez, et Mariano Fernández-López. 2012. «The NeOn Methodology for Ontology Engineering». In *Ontology Engineering in a Networked World*, 9-34. Springer Berlin Heidelberg.

Tamma, Valentina, et Trevor Bench-Capon. 2001. «A Conceptual Model to Facilitate Knowledge Sharing in Multi-Agent Systems».

Uschold, Mike, et Michael Gruninger. 1996. «Ontologies: Principles, Methods and Applications». *The Knowledge Engineering Review* 11 (02): 93.

Wardhana, Helna, Ahmad Ashari, et Anny Sari. 2018. «Review of Ontology Evolution Process». *International Journal of Computer Applications* 179 (mars): 26-33.

Zablith, Fouad, Grigoris Antoniou, Mathieu d'Aquin, Giorgos Flouris, Haridimos Kondylakis, Enrico Motta, Dimitris Plexousakis, et Marta Sabou. 2015. «Ontology evolution: A process-centric survey». *The Knowledge Engineering Review* 30: 45-75.

Résumé

L'évolution d'ontologie fait partie des activités du cycle de vie d'une ontologie et représente un réel challenge dans le domaine de l'ingénierie des ontologies. En raison de la variété des conséquences causées par les changements que peut subir une ontologie (changements des exigences, modification des besoins, évolution des connaissances modélisées, etc.), une attention particulière doit être apportée au processus d'évolution d'ontologie afin de garantir sa cohérence et sa consistance. La stratégie adoptée dépend de l'ampleur des modifications à apporter sur l'ontologie conçue. L'objectif de cette étude consiste à présenter les étapes menant à l'évolution d'une ontologie modulaire. Nous identifions ainsi les connaissances permettant d'appliquer les changements nécessaires à l'évolution des modules existants puis au développement de nouveaux modules. Nous présentons, à travers cette évolution, des exemples concrets d'inférences obtenues pour personnaliser la recommandation d'aliments et d'activités selon les données collectées sur la personne âgée.

The Harmonization of Terms and Concepts in Function of the Standardization of Computer Terminology in the Albanian Language

Anila Çepani*, Adelina Çerpja**

* Faculty of History and Philology, University of Tirana, Rruga e Elbasanit 2, Tirana, Albania, https://www.fhf.edu.al/
anila.cepani@unitir.edu.al
** Institute of Linguistic and Literature, Academy of Albanian Studies, Sheshi "Nënë Tereza", nr. 3, Tirana, Albania, http://asa.edu.al/site/igjl/
adelina.cerpja@asa.edu.al

Abstract. It is a known fact nowadays that humanity is experiencing the fourth technological revolution, that is bringing a fast and strong social and economic impact on a global scale, which could not but be accompanied by the creation of a new specialized language that tries to follow this uninterrupted development and change. Therefore, computer technology terminology is one of the most problematic issues of specialized languages in various fields of knowledge.

As a result of the successive advanced versions of electronic devices, as well as the constant updates of various software and applications, a continuous enrichment of the relevant terminology is required, so that the terms adapt to these changes.

Despite several years of work on several plans for the creation and standardization of computer terminology in the Albanian language, it seems that it has not yet taken the right place, occasionally leaving space for the use of relevant terms in English. One of the main factors that have influenced this use is the lack of a good part of the cases and terms unified in the Albanian language. In the existing dictionaries of computer technology terms, there are uses of different variants for a term, and this fluctuation is noticed even in the same dictionary. This lack of standardization of terms makes the user feel insecure, resulting in the rapid 'embrace' of the term in English.

In this paper we aim to present the state of use of computer terminology in the Albanian language, to make an analysis of the causes of the lack of unification of these terms and then, through a comparative look, to dwell on some concrete examples of computer terms which appear with various choices in the dictionaries, followed by some modest thoughts on how to work towards updating and standardizing computer terminology in the Albanian language.

1. Introduction

Unifying the terminology of different fields of knowledge poses a permanent challenge, because defining a term for a concept requires a deep knowledge of the field and permanent collaboration between linguists and relevant specialists. Unlike the terminology of other fields of science, such as mathematics, physics, chemistry, etc., which is generally standardized and widely accepted within the scientific community, for some other fields of knowledge, especially relatively new and intensively evolving ones, like computer technology, the issue of unification of terms is more complicated. The revolution that new information and communication technologies (including e-mail, Internet, all user-computer interaction, information digitization, satellite communication, etc.) have brought to our society has had a tremendous impact on language, thus creating a new specialized language, which must be flexible enough to adapt successfully to the industry's ongoing innovation. Given the intensity of the releases of different versions of equipment, as well as the successive updates of software and applications, great and continuous work with the relevant terminology is required, which must go hand in hand with these changes. Therefore, nowadays it is a fact that the terminology of computer technology and information is one of the most problematic issues of specialized languages for different fields of knowledge.

2. Methods and materials

This paper is the result of several years of work of the authors in the project for the localization in Albanian of Windows and Office's several versions. The process of localization was long and difficult and required cooperation between technical and linguistic experts. This project would be of great importance, because it was the first experience of localizing and using a massive software in the Albanian language. It would also serve as a basis for the subsequent localization of other software to which a higher technical and linguistic level is evident.

It should be noted that localization is not an ordinary translation process. Localization is the process of adapting a product, in our context a software program, to a specific locale, i.e., to its language, standards and cultural norms as well as to the needs and expectations of a specific target market. A properly localized product also must meet all the legal requirements in force in the user's region. According to one of the main companies of localization[1], the standard localization process includes the following basic steps:

- Analysis of the material received, and evaluation of the tools and resources required for localization
- Cultural, technical, and linguistic assessment
- Creation and maintenance of terminology glossaries
- Translation to the target language

The most important phase of localization, at the same time and the first step of the work, was the preparation of the preliminary dictionary with a certain number of terms in the Albanian language, which helps to use a more unified terminology during localization. Each term of the English dictionary, which was the source dictionary, was reviewed with the aim of adapting it to the Albanian language. As has happened and continues to happen in other languages, localization is undertaken mainly from English, because many software are developed in the USA and the software in other countries are originally mainly developed in English. It is difficult to translate or adapt terms from one language to another, in this case from English to Albanian, because there are some terms which are impossible to translate, but, even if they are translated, they lose their meaning.

During the work for localization and especially during the compilation of the dictionary, various problems were encountered, which can be grouped into several categories: *Finding the appropriate term; Accurate adaptation of the defining word to terms with more than one word; Difficulty in determining which part of speech the term belongs to; The ambiguity of a term,* etc.

Descriptive and comparative methods are used to bring to attention the contribution of various scholars in relation to this issue, focusing on the study of the similarities and differences encountered between them. This comparative look has helped us to highlight the problems encountered in the field of computer terminology in the Albanian language.

Quantitative and qualitative methods have been used to extract and analyze concrete data in terms of computer technology: current terminological dictionaries of this field as well as any related work have been included in the concordance program enabling the parallel analysis and comparison of these terms by different authors.

[1] https://www.trados.com/solutions/software-localization/

3. The creation of a database for computer terms in the Albanian language

Given the dense use of these terms because of the massive use of diverse technological equipment, there is a natural need for elaboration and unification of terminology of this specific field in the Albanian language. Despite the many difficulties, it has been continuously worked on several plans for the creation, adaptation, and standardization of terms in this field and as a result, a good database of Albanian terms has been created. These results have been achieved through work in these areas:

- Compilation of several dictionaries for computer terminology;
- Localization of Microsoft programs, specifically Windows, as the operating system most used among Albanians, as well as variants of the Office package, programs, and other applications suitable for computers or mobile phones;
- The trend to include the most used computer terms in explanatory (monolingual) dictionaries[2] of Albanian.
- Compilation of school and university textbooks, for information technology, in which terms are generally used in the Albanian language.

During the process of elaborating the terminology of a certain field, specifically that of computing, researchers aim to create terms as clear and transparent as possible, to avoid ambiguity. These connections, reflected in the synonyms and homonyms that often aggravate this terminology, are found in use in various guides and applications, in school and university textbooks, in explanatory dictionaries, as well as bilingual or disaggregated dictionaries, etc. Special attention should be paid not only to the content of various scientific publications, textbooks, and university textbooks but also to the form, terms used, the use of which naturally affects their content.

[2] The explanatory dictionary of Albanian 2006 includes about 44 terms in this field: *disketë, faqos*, *faqosës*, *faqoset*, *faqosje*, *faqosur*, *formatim*, *formatohet*, *formatoj*, *formatuar*, *frontespic*, *internet, intranet, klikim*, *klikoj*, *kompaktdisk*, *kompjuter, kompjuterik, kompjuterizoj, kursor*, *laser, laserik, maus*, *mikrokompjuter, mikroprocesor, miush*, *monitor, monitorim, monitoroj, printer, printer, printim, printoj, printuar, programues*, *radhit*, *radhitës*), *rul*, *skaner*, *skanim*, *skanohem*, *skanoj*, *skanuar*, *terminal*. In the dictionary of 2002, we did not find any term of computer terminology, although at the time of compiling this dictionary the use of computer and internet had a massive spread; not even the term computer, with which this dictionary is written, is included.

As in any terminological dictionary, computer terminology uses elements from the general lexicon which are used in a more specific sense, depending on the context in which they are encountered. Being an active part of the general lexicon, with all the semantic expansion they undergo, these terms are more quickly accepted and easily used. Such terms are: (accept) *pranoj*, (action) *veprim*, (bind) *lidh*, (channel) *kanal*, (choose) *zgjedh*, (clear) *pastroj*, (close) *mbyll*, (delete) *heq*, (error) *gabim*, (field) *fushë*, (folder) *dosje*, (port) *portë* etc.

A significant place in computer terminology is occupied by terms borrowed from the terminology of other fields, mainly technical, such as: (clone) *klon*, (diacritic) *diakritik*, (fraction) *thyesë*, (frame) *kuadër*, (gradient) *gradient*, (index) *indeks*, (matrix) *matricë*, (monitor) *monitor*, (parse) *analizoj*, (screen) *ekran*, etc.

One of the common features of terminological dictionaries is to give the concept with a single word, so even in these computer terminology dictionaries, one-word terms predominate: (developer) *zhvillues*, (eject) *nxjerr*, (dock) *stacionoj*, (erase) *fshij*, (frame) *kuadër*, (install) *instaloj*, (interface) *ndërfaqe*, (node) *nyjë*, (password) *fjalëkalim,* etc.

However, there are terms for which it has been impossible to be expressed with a single word in Albanian, such as: (foreground) *plan i parë*, (hyphen) *vijë ndarëse*, (idle) *gjendje e qetë*, (benchmark) *testues i performancës*, (bookmark) *shenja e referimit*, (breakpoint) *pika e ndalimit,* etc.

Often in these cases, the one-word terms have played the role of the governing element (head), but sometimes they also appear as determinant words:

English	Albanian
application	***aplikacion***
application binary interface	*ndërfaqe binare e* ***aplikacionit***
application developer	*zhvillues i* ***aplikacioneve***
application development environment	*rrethinë për zhvillim të* ***aplikacionit***
application development system	*sistem për zhvillim të* ***aplikacionit***
application file	*skedar i* ***aplikacionit***
application framework	*struktura e* ***aplikacionit***
application icon	*ikonë e* ***aplikacionit***
application layer	*shtresë e* ***aplikacionit***
application package	*paketë softuerike e* ***aplikacionit***

English	Albanian
application program interface	*ndërfaqe e programit të* ***aplikacionit***
application program	*program i* ***aplikacionit***
application programming interface	*ndërfaqja e programimit të* ***aplikacionit***
application protocols	*protokolle të* ***aplikacionit***
application server	*server i* ***aplikacionit***
application software	*softuer i* ***aplikacionit***
application shortcut key	*tast i shkurtores së* ***aplikacionit***
application window	*dritare e* ***aplikacionit***
applications programmer	*programues i* ***aplikacioneve***.

English	Albanian
access	***qasje***
access name	*emri i* ***qasjes***
access number	*numri i* ***qasjes***
access path	*shtegu i* ***qasjes***
access right	*e drejta e* ***qasjes***
access server	*serveri i* ***qasjes***
access time	*koha e* ***qasjes***
access type	*lloji i* ***qasjes***

English	Albanian
filter	***filtroj, filtrim,*** *filtër*
filter area	*zona e* ***filtrit***
filter by form	***filtrim*** *sipas formës*
filter by selection	***filtrim*** *sipas përzgjedhjes*
filter excluding selection	***filtrim*** *duke përjashtuar përzgjedhjen*
filter field	*fusha e* ***filtrit***

These one-word terms are regularly used used as source words for building new word, such as : ***printoj***, *printer*, *printim*, *i printuar* ; ***përditësoj***, *përditësim*, *i përditësuar* ; ***bashkëngjit***, *bashkëngjitje*, *i bashkëngjitur*, etc.

4. Problematics of non-normative use of computer terms

Despite several years of work on different plans for the creation and standardization of computer terminology in the Albanian language, it seems that it has not yet taken the right place, occasionally leaving space for the use of relevant terms in English. One of the main factors that have influenced this use is the lack of unified terms in most cases in the Albanian language.

In the existing dictionaries of computer technology terms, there are uses of different variants for a term, and this fluctuation is noticed even within the same dictionary. This lack of standardization of terms makes the user feel insecure, resulting in the rapid 'embrace' of the term in English.

In most cases, non-normative uses occupy a large percentage. Often several variants are used for a term, sometimes phonetic variants of Albanian terms, sometimes different variants of terms in foreign languages, mainly in English:

(accept) the normative term *pranoj* but also *akseptoj / akceptoj ;*
(attachment) *bashkëngjitje* but also *ataçment / atashment ;*
(cancel) *anuloj* but also *kanceloj / bëj cancel ;*
(database) *bazë e të dhënave* but also *databeiz / databeis / database databazë / të dhënat bazë ;*
(file) *skedar* but also *dokument / fajll ;*
(option) *opsion* but also *opcion / obcion ;*
(provider) *furnizues* but also *provajder / provider ;*
(router) *rrugëzues* but also *ruter / rrugëzim ;*
(web browser) *shfletues uebi* but also *veb-brauzer / web shfletuesi / ueb browser ;*
(shortcut) *shkurtore*, *shkurtesë*, *shortkat*, *shortcut*, etc.

The presence of different variants for the same term is indicative of non-inclusion in the proper degree of normative terms. The compilation and the use of dictionaries of different types such as explanatory, monolingual, or multilingual, dictionaries of different sizes are very important, so that every user, whether ordinary or specialist in a certain field, can use these terms properly.

Attention should be paid not only to the content of various scientific publications, textbooks, and university textbooks but also to the terms used in them, which would increase the awareness of users and lead them gradually toward the use of a unified terminology in information technology.

We have made an observation about the state of use of these terms either in various forums or 'online' sites or in various school and university publications. Leaving aside the cases of lack of use of unified Albanian terms (it happens that within the same text there are different variants of the term in Albanian), we have concluded that there are three main problems with the use of computer terms :

1. Unnecessary use of terms in English in their original form (without inflectional endings or with endings, but written adjacent, without a hyphen) even though there is an equivalent term in the Albanian language for them. We are listing below some of them, putting in parentheses the Albanian version that could have been used : *adapter* (përshtatës), *akses* (qasje), *update* (përditësoj), *bookmark* (shenja e referimit), *booting* (procesi i fillimit), *browser* (shfletues), *checkbox* (kuti zgjedhjeje), *clipboard* (kujtesa e fragmenteve), *clipart* (fragment artistik), *database* (bazë të dhënash), *driver* (drejtues), *file* (skedar), *folder* (dosje), *firewall* (mur mbrojtës), *interface* (ndërfaqe), *keyboard* (tastierë), *compiler* (përpilues), *compatibile* (i përshtatshëm), *link* (lidhje), *manager* (menaxhues), *network* (rrjet), *navigation* (lundrim), *password* (fjalëkalim), *pointer* (tregues), *portable* (i lëvizshëm), *procesing* (përpunim), *rang* (diapazon), *site* (sajt), *scroll* (lëviz), *slide* (pamje rrëshqitëse), *slot* (fole shtesë), *subfolder* (nëndosje), *starting* (nisja), *shorcut* (shkurtore), *startupi* (nisja), *template* (model), *taskbar* (shiriti i detyrave), *toolbar* (shiriti i veglave), *web* (ueb), *website* (sajt uebi), etc.
2. Unnecessary use of terms in English that are written according to the pronunciation and phonetic system of the Albanian language, while the term in Albanian is clear and fully covers the meaning of the original term. In these cases, there are different forms, which are written according to the different pronunciation of the English term : *apgreid* (përmirësoj); *apload* (ngarkoj); *bekap* (rezervoj); *bekgraund* (sfond); *brauser* (shfletues); *drajva*, *drajvera* (drejtues); *enkapsulon* (vë në kapsulë); *fajl*, *fajll* (skedar); *follder* (dosje); *futer* (fundi i faqes); *imeil*, *imejl* (postë elektronike); *interfejs*, *interfes* (ndërfaqe); *kejbord* (tastierë); *kompajler* (përpilues); *kompatibil*, *kompatëbël* (i përputhshëm); *kraker* (thyerës); *lokacion* (vendndodhje); *matherbord* (bord qendror); *particion* (ndarje); *pasuord*, *pasvord* (fjalëkalim); *ruter* (rrugëzues); *slaid*, *slajd*, *sllajd* (pamje rrëshqitëse); *stringje* (vargje), etc.
3. Terms created through Albanian suffixes, derived from English stems, pronounced and written according to the Albanian phonetic system : *adaptimin* (përshtatjen) from *adapt*, *aksesimin* (qasjen) from *access*, *butimit* (nisjes) from *boot* ; *çekimin* (zgjedhjen) from check, *daunloadoj*

ose, *danlodoj* (shkarkoj) from *download*, *editimi* (redaktimi) from *edit*, *konektimin* (lidhjen), *ristartim* (rinisje) from *restart*, *selektoj* (përzgjedh) from *select*, *selektim* (përzgjedhje) from *selection*, *startoj* (nis) from *start*, *startim* (nisje) from *start*, *updetuar*, *apdetuar* ose *apdejtuar* (përditësuar) from *update*, etc.

5. Categories of use of terms in the source language

In the computer terminology of the researched sources, although there are several Albanian terms, equivalent to those in English, there is a strong tendency to use terms in the source language. According to the statistical observation of terms taken from texts and conversations in forums, more specifically, based on the frequency of usage, we divide into three main categories the English borrowings present in the Albanian computer terminology of school and university textbooks, as well as in conversations in various forums:

a) terms used on the same frequency as the Albanian ones: e.g.: *hard disk / disk i ngurtë*; *monitor / ekran*; *link / lidhje*; *network / rrjet*; *browser / shfletues,* etc.

b) terms that are used more often than the Albanian equivalents: *lokacion*, *drajver* (*driver*), *stringje* (strings), *slajd* (*slide*), etc.

c) terms used exclusively in the source language because there is no equivalent in Albanian: e.g.: *server*, *banners*, *cookie*, *skaner*, *routers*, etc.

6. Comparative observation of terms in different terminological dictionaries

To have a clearer picture of the standardization of computer terminology in the Albanian language, we have made a comparative observation of terms in three dictionaries[3] of informatics in the Albanian language, which constitute a valuable contribution to the creation and standardization of this terminology.

[3] *Fjalor i informatikës* (*anglisht – shqip*; *shqip – anglisht*) [Dictionary of Informatics (English - Albanian; Albanian - English)], Kosovo Academy of Sciences and Arts, Nebi Caka, Agni Dika, Seb Rodiqi, editor Rexhep Ismajli, Pristina, 2005; *Fjalor enciklopedik nga teknologjia e informacionit*, *anglisht – shqip – gjermanisht - frëngjisht* [Encyclopedic dictionary from information technology, English - Albanian - German – French], Mehmeti, A. Sh., Pristina, 2006; *Fjalor i termave të informatikës* (anglisht-shqip-anglisht) [Dictionary of informatics terms (English-Albanian-English)], L. Shishani, A. Çerpja, & A. Çepani, Pristina, 2010.

However, it is noted that in addition to common terms, the authors have given different solutions for some English terms, even within the same dictionary. Based on this comparative observation, several phenomena are noticed:

- The same term in Albanian is used instead of the source term in English. This is encountered both in terms derived from words with a general meaning, and in specialized terms, being a good indicator for the unification of this terminology: (attach) *bashkëngjit*, (browser) *shfletues*, (backup) *kopje rezervë*, (end-user) *përdorues fundor*, (router) *rrugëzues*, (data compression) *ngjeshje e të dhënave*, (update) *përditësoj,* etc.
- Different solutions are given for a considerable number of source terms. This phenomenon adds to the variety of terms, especially when it comes to terms that have a high frequency of use and which serve as heads for the construction of multi-word terms: (access) *qasje*, *hyrje*, *kam akses në*, *qasem*; (array) *varg*, *matricë*, *fushë*, *vektor*, *tabelë*, *sërë*; (add-on) *modul*, *pajisje*, *pllakë*, *shtesë*; (booting) *procesi i fillimit*, *inicim*, *ngritje e sistemit / fillim*; (chat) *muhabet*, *kuvendim*, *bisedë*; (connector) *konektor*, *bashkues*, *lidhëzor*; (crash) *prishje / rënie e sistemit*, *ndërprerje aksidentale*, etc.
- The variety of terms used is also reflected in the multiword terms, in which these are used as headings or determining words: (backup file) *skedar rezervë*, *kopje e skedarit*, fajl rezervë; (acceptable use policy) *politikë e pranueshme e përdorimit*, *politikë e pranueshme e shfrytëzimit*, *rregullorja e përdorimit të pranueshëm*; (access code) *kod i hyrjes*, *kodi i qasjes*, *kod i aksesit*; (access control) *kontroll i hyrjes*, *kontrolli i qasjes*, *kontrolli i aksesit*; (access point) *pika e hyrjes*, *pika e qasjes*, *pika e aksesit*; (boot sector) *sektor për procesin e fillimit të sistemit*, *sektor për inicimin e sistemit*, *sektor për ngritje të sistemit*, *sektor për fillimin*; (chat room) *dhomë e muhabetit*, *dhomë bisede,* etc.
- Often the authors of the same dictionary, to precede each context, have given several variants for the same concept separated by commas from each other. We think that in some cases it has been unnecessary and has confused the user, especially when these terms appear as an integral part of multiword terms. In these cases the users choose different variants every time they needed to use that term, for example: (interface) *ndërfaqoj / lidh / bashkoj;* (interface) *ndërfaqe / interfejs / pajisje ndërmjetësuese;* (interoperability) *ndërfunksionalitet / ndërveprueshmëri / pajtueshmëri*; (iterative array) *matricë përsëritëse / matricë iterative;* (jaggies) *dhëmbëzimet / shkallëzimet*; (jumper) *urëzues / shkurtlidhës*

/ kapërcyes / urë / urë lidhëse ; (memory cache) *memorie e gatshme / memorie e fshehtë ;* (navigation) *navigim / shfletim / orientim* ; (nickname) *pseudonim / nofkë / emër alternativ ;* (node) *nyjë / pikëdegëzim / pikëbashkim ;* (online processing) *përpunim në-linjë / përpunim i drejtpërdrejtë me kompjuter / përpunim në kohë reale ;* (phone connector) *lidhëzor telefoni / konektor telefoni.*

- There are also cases of lack of continuity in the use of the main term and the other term derived from that, such as *attach = bashkëngjit*, but *attachment = shtojcë*, *compatible = pajtueshëm* (*i*), but *compatibility = përputhshmëri*, etc.
- Fluctuation in some terms with two or more words comes as a result of incorrect selection of the governing head and the determining words, resulting in two different terms, as in the case of : *object-oriented graphics - grafikë e orientuar në objekte - grafik me objekt të orientuar* ; *absolute cell reference - referencë absolute e qelizës -referencë e qelizës absolute ; central processing unit - njësia qendrore e përpunimit - njësia e përpunimit qendror ; command prompt window - dritarja e komandave të menjëhershme - dritarja e menjëhershme e komandave,* etc.

7. Influencing factors of the problematics encountered

We think that the problems encountered in our sources, such as in various media or ‘online’ forums, in textbooks or publications of the university and pre-university system, in terminological dictionaries, and in other situations in which terminology is used, are directly or indirectly consequences of several factors :

1. The very important role of the rapid industrial evolution of this sector and the constant need for new terms.
2. There was a lack of organized work for the unification of these terms in different computer dictionaries, making the user not feel safe and quickly use the term in English or use different variants in the Albanian language rather than a standardized term. Despite the localization of some versions of Windows and Office packages, the use of these programs in Albanian was not massive, which would have led to the implantation of Albanian terms, so they would have seemed more natural to everyone even in other contexts.
3. The frequent launch of new versions of various programs, albeit with minimal changes, would always require their localization, but this is

very difficult to achieve, especially when the previously undertaken initiatives for the localization of various programs in Albanian are discontinued. Consequently, those users who have previously used the Albanian versions will return to the English versions in their everyday usage. Also, the Albanian versions of some applications for smartphones are made by non-professionals, who have not always used the appropriate terminology in Albanian, leading to the disuse of unified terms in Albanian and the use of the English ones.

4. Often computer science specialists, especially programmers, show indifference to Albanian terms, as they are familiar with those in English and their experiences make them naturally use English terms in their work during programming.

8. Suggestions for improving the situation

For dictionaries that will be prepared in the future, whether terminological or explanatory dictionaries, it is recommended for authors to consult with existing computer dictionaries, as well as specialists in the relevant fields, before including computer terms in explanatory dictionaries. According to the researcher Jani Thomai in his paper on anglicisms in the explanatory dictionaries of the Albanian language, 50 anglicisms of new fields of development were added in the electronic version of the Albanian language dictionary of 2012, borrowed mainly during the last two decades, among which he mentions informatics terms, such as *akses*, *bajt*, *bold*, *databeiz*, *desktop*, *fail*, *feisbuk*, *haker*, *imeil*, *insert*, *kilobajt*, *pasuord*, *selektoj* (*selektim*), etc. (Thomai, 2014 :231-232). We think that for most part, the use of terms in the Albanian language has already been consolidated and they should also be included in explanatory dictionaries, which have a greater impact on ordinary users.

Supposing that a dictionary of the large type will have to include many other recent anglicisms that are not yet included in the explanatory dictionaries of the Albanian language, J. Thomai lists the following computer technology terms : *apdejtoj* (*apdejtim*) (*also fig.*), *bekgraund*, *gigabajt*, *harddisk*, *harduer*, *kompaktdisk*, *laptop*, *modem*, *monitor*, *selektoj*, *servër*, *softuer*, *çat* (*çatoj*, *çatim*), *displej*, *drajv*, *insert*, *mikroçip*, *peixhmekër*, *uebsajt*, etc. We think that the inclusion of most of these terms or the suggestion to become part of future dictionaries should be viewed with caution, because for most of these terms very good solutions are given in the existing dictionaries of computer terms, a part of which have massive use : *qasje* for *akses*, *skedar /*

fail, *ndërfus / insert*, *fjalëkalim / pasuord*, *përzgjedhje / selektim*, *sfond / bekgraund*, *ekran / monitor*, *përditësoj / apdetoj*, *afishoj / displej,* etc.

9. Conclusion

Computer terminology has already become an important part of the technical-scientific terminology of Albanian. A continuous work has been done for the Albanian language in recent years using its possibilities and means of expression, always keeping in mind the preservation of the scientific accuracy of the concept. But we are aware that it is necessary to continue working, and here are some suggestions to follow.

Compiled dictionaries of computer terms and others that will follow constitute important steps in the field of informatics terms, serving as a good basis for further enrichment and improvement with new terms. Given that this new information technology is developing at a gigantic pace and that the number of new terms is increasing at the same pace, more fruitful cooperation of specialists in both fields is needed to achieve a standardization of this terminology in Albanian. The inclusion of terms without consulting the vocabularies of informatics terms would undo that hard work and would stress the problem in the terminology of this field.

Given the totality of problems that characterize the use of information technology terms in the Albanian language, we think that more work should be done to continue the standardization of these terms by following clear scientific principles and criteria, respecting the tendency to avoid where it is possible foreign terms and to use the lexical resource, as well as different models of word formation in the Albanian language.

We also value the compilation of dictionaries with terms that have a higher frequency of use, dictionaries which, being more practical than other dictionaries, can be used more widely.

But today is not the time to use only traditional, printed dictionaries. Today is the time to use electronic versions of dictionaries, whether installed or on the web. This work has already started, but it is very important for Albanian to continue at a faster pace in this direction.

Another task that is posed and that directly affects the unification and use of these terms is the localization of computer terms, especially in Office packages or in major operating systems, such as Microsoft Windows, which are widely used by Albanian (according to most data latest for 2020 is the most used among Albanians in Albania (https://gs.statcounter.com/os-market-share/desktop/albania) and Kosovo (https://gs.statcounter.com/os-market-share/desktop/kosovo). This would not only bring great convenience to the user but would affect the faster and wider spread of computer terms in the Albanian language, as well as their perfection and continuous updating.

We consider it of special importance to undertake a project for the creation of a 'bank' of computer technology terms in the Albanian language, in Albania, Kosovo, and Northern Macedonia, to reach a unified terminology, which should be used in various pre-university and university publications.

This would increase the awareness of users and gradually the use of a unified terminology in this important area of economic and social development, but also political, at a time when all three countries aspire to the EU and aim to meet its standards.

Reference

Caka, N. (2007). Standardizimi i terminologjisë së informatikës. *Gjuha jonë*, 1-4, p. 50-53.

Caka, N. (2005) "Terminologjia e informatikës në gjuhën shqipe dhe standardizimi i saj në fjalorë", *Leksikografia shqipe – Trashëgimi dhe perspektivë*, Tiranë, f. 239-249.

Concise Oxford English Dictionary (2001). Tenth edition on CD-ROM, version 1.1.

Çepani, A. (2018). "Ndikimi i anglishtes në terminologjinë kompjuterike shqipe dhe gjendja sot", Studime albanologjike, 2018/1, Viti XXI, Universiteti i Tiranës, Fakulteti i Historisë dhe i Filologjisë, Tiranë, f. 207-217.

Çerpja, A. & Çepani, A. (2014) "Computer Terminology in Albanian versus Other Languages and the Benefits of Its Codification", *Journal of Education and Practice*, Amerikë, f. 42-47.

Çepani, A. & Çerpja, A. (2018). "Globale Entwicklungen der Computerterminologie. Das Albanische im Vergleich". in- *The Potentiality of Pluricentrism. Albanian Case Studies and Beyond.* Albanische Forschungen, Band 41. Edited by Lumnije Jusufi. Harrassowitz Verlag. Wiesbaden. ISSN 0568-8957. ISBN 978-3-447-11160-7, f. 185-198.

Çepani, A. & Çerpja, A. (2007). "Probleme të lokalizimit të termave të teknologjisë informative në shqip", *Seminari I Ndërkombëtar i Albanologjisë*, Tetovë – Ohër, f. 192-203.

Çepani, A. & Çerpja, A. (2010). “Shqipja dhe terminologjia kompjuterike”, Seminari XXIX Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare, Prishtinë, f. 275-282.

Dictionary of Computer and Internet terms (2002). seventh edition, Columbia Press, Coypiright by Columbia University Press, New York.

Drejtshkrimi i gjuhës shqipe (1973). Akademia e Shkencave, Instituti i Gjuhësisë dhe i Letërsisë, Tiranë.

Duro, A. (2001). *Terminologjia si sistem*, “Panteon”, Tiranë, 147 p.

Fjalor i gjuhës shqipe (2006). Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë, Tiranë, 1252 f.

Fjalor i informatikës (*anglisht – shqip*; *shqip – anglisht*) (2005). Akademia e Shkencave dhe e Arteve e Kosovës, Nebi Caka, Agni Dika, Seb Rodiqi, redaktor Rexhep Ismajli, Prishtinë.

Fjalor i shqipes së sotme (2002). botim i dytë i ripunuar, Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë, Tiranë.

Fjalor i termave themelorë të mekanikës (2002). Redaktor Agron Duro. Akademia e Shkencave e Shqipërisë. Instituti i Gjuhësisë dhe i Letërsisë, Shtëpia Botuese “Shkenca”, Tiranë, 1224 f.

Hoti, I. (2001). *Fjalori enciklopedik anglisht – italisht – shqip për bankën*, *biznesin*, *bursën*, *ekonominë*,... *internetin*..., Durrës. http://www.webopedia.com/

Hysa, R. (1998). *Fjalor anglisht–shqip*, Tiranë.

Mehmeti, A. Sh. (2006). *Fjalor enciklopedik nga teknologjia e informacionit*, anglisht – shqip – gjermanisht - frëngjisht, Prishtinë.

Microsoft Computer Dictionary (2002). fifth edition.

Për pastërtinë e gjuhës shqipe (1998). Akademia e Shkencave e Shqipërisë, Tiranë.

Roncaglia, G. (2003). *Il topo scannato. Italiano e terminologia informatica*, Convegno internazionale “Lingua italiana e scienze“, Accademia della Crusca, Firenze. http://www.merzweb.com/testi/saggi/italiano_e_terminologia_informatica.htm

Shishani, L., Çerpja, A. & Çepani, A. (2010). *Fjalor i termave të informatikës* (anglisht-shqip-anglisht), Prishtinë.

The IBM Dictionary of Computing (1994). (On disk) based on the printed IBM Dictionary of Computing published by McGraw-Hill.

Thomai, J. (2014). Anglicizmat në fjalorët shpjegues të gjuhës shqipe [Anglicisms in the explanatory dictionaries of the Albanian language], *Studime* (Journal of Philological Studies) 20/2013, Academy of Sciences and Arts of Kosovo, Prishtina, 2014, p. 231-232.

Veselaj, N. (2000). Rreth terminologjisë kompjuterike në shqipe. *Gjuha shqipe*, 1-2, p. 35-44.

https://gs.statcounter.com/os-market-share/desktop/albania

https://gs.statcounter.com/os-market-share/desktop/kosovo

Websites for texts

https://botaedituris.files.wordpress.com/2012/09/kompjuteri-ndertimi-dhe-funksionet-_hardware_gati.pdf

http://klubiteknikmekatronikagjonserrec.weebly.com/uploads/1/1/6/4/11649867/libri-i-msuesit-informatika-.pdf

http://klubiteknikmekatronikagjonserrec.weebly.com/uploads/1/1/6/4/11649867/informatika_.13.pdf

http://www.ideart.al/wp-content/uploads/2013/08/Liber-Mesuesi-Inf-7-8-9.pdf

http://www.ideart.al/wp-content/uploads/2013/08/L-M-TIK-10-Berthame.pdf

http://digitalbook.botimepegi.al/pluginfile.php/12526/mod_page/content/39/TIK%206.pdf

http://digitalbook.botimepegi.al/pluginfile.php/12526/mod_page/content/37/TIK%2011.pdf

http://www.elve.al/images/libra-mesuesi/Liber%20mesuesi%20informatik%2011.pdf

http://www.botimepegi.al/phocadownload/Mesuesi2016/Klasa7/informatika%207%20-%20liber%20mesuesi.pdf

http://www.forumishqiptar.com/threads/55692-Ndryshimi-i-menys%C3%AB-s%C3%AB-boot-it-%28linux-windows%29

chrome-extension://ohfgljdgelakfkefopgklcohadegdpjf/https://shkollaime.al/pluginfile.php/29/mod_page/content/74/TIK%207.pdf

https://forum.itshqip.com/threads/problem-me-bootim-ne-windows-7.40988/

https://programe.albanianforum.net/f30-informatika

chrome-extension://ohfgljdgelakfkefopgklcohadegdpjf/https://shkollaime.al/pluginfile.php/29/mod_page/content/74/TIK%2010.pdf

Résumé

Aujourd’hui l’humanité connaît la quatrième révolution technologique, très rapide et avec un fort impact social et économique au niveau mondial, qui ne pouvait que s’accompagner de la création d’un nouveau langage spécialisé qui tente de suivre pas à pas ce développement et ce changement continuel.

C’est pourquoi la terminologie informatique est l’une des questions linguistiques les plus problématiques des langages spécialisés dans divers domaines de la connaissance.

En raison des versions de plus en plus avancées des appareils électroniques, ainsi que de la mise à jour constante de divers logiciels et applications, un enrichissement continu de la terminologie pertinente est nécessaire, de sorte que les termes répondent en parallèle à ces changements.

Malgré plusieurs années de travail sur certains projets de création et de standardisation de la terminologie informatique en albanais, il semble qu’elle n’ait pas encore pris la bonne place, laissant parfois la place à l’utilisation de termes pertinents en anglais. L’un des principaux facteurs qui a influencé cette utilisation est le manque dans une bonne partie de cas et de termes unifiés dans la langue albanaise. Dans les dictionnaires existants de termes informatiques, il existe des utilisations de différentes variantes pour un terme, et même cette fluctuation est remarquée même au sein du même dictionnaire. Ce manque de standardisation des termes rend l’utilisateur peu sûr, ce qui entraîne une “adoption” rapide du terme en anglais.

Dans cet article, nous cherchons à présenter l’état d’usage de la terminologie informatique en albanais, à faire une analyse des causes du manque d’unification de ces termes et ensuite, à travers un regard comparatif, à s’attarder sur quelques exemples concrets des termes informatiques pour lesquels les dictionnaires de ce domaine divers choix ont été faits, suivis de quelques réflexions modestes sur la manière de travailler à la mise à jour et à la normalisation de la terminologie informatique en albanais.

Keywords

Term, terminology, computer, term unification, standardization.

Extending TBX2RDF

Thierry Declerck*/**, Patricia Martín Chozas***, Tom Winter****,
Tanja Wissik**

*DFKI GmbH, Saarland Informatics Campus
Multilinguality and Language Technology Lab
Stuhlsatzenhausweg, 3
D-66123 Saarbrücken, Germany
declerck@dfki.de
https://www.dfki.de/~declerck /
**Austrian Academy of Sciences
Austrian Centre for Digital Humanities and Cultural Heritage
Sonnenfelsgasse, 19
A-1010 Vienna, Austria
tanja.wissik@oeaw.ac.at
https://www.oeaw.ac.at/acdh/team/current-team/tanja-wissik
***Universidad Politécnica de Madrid
Ontology Engineering Group
Campus de Montegancedo, s/n. ETSIINF, 28660 Boadilla del Monte, Madrid, Spain
pmchozas@fi.upm.es
****Deutsche Bahn AG
Sprachenmanagement (IBS)
Gallusanlage, 8
D-60329 Frankfurt a. Main, Germany
Tom.Winter@deutschebahn.com

Abstract. Following the goal of transforming two multilingual terminologies to a native graph representation format, we propose an extension of an already existing initiative tackling this matter, named as TBX2RDF, which leads to the design of a new RDF-based modelling of traditional terminological data. The original multilingual terminological data handled in this work come from the terminology of the Deutsche Bahn AG and the Interactive Terminology for Europe (IATE).

1. Introduction

We present in this paper results from our work which consists in porting two multilingual terminologies from their TBX[1] representation onto an RDF[2] format. While the purpose of this exercise is not to change anything at the level of the content of the original terminologies, their modelling in a native graph-based representation offers the possibility of interlinking and merging them with other resources, being in the realm of terminologies or other types of resources, like for example detailed lexicographic resources, knowledge bases and even ontologies.

Within this context, we came into several modelling suggestions that go beyond the TBX2RDF converter[3], an existing application to map TBX data to RDF vocabularies, which we also present in this paper. This eventually led to the design of a new RDF-based modelling approach for terminologies, which is now being discussed within the Ontolex W3C Community Group[4]. Through the following sections, we present the two multilingual terminological resources that are the core of our modelling work. In Section 4, we describe the LIDER project[5] in which the TBX2RDF service was developed, alongside with the resulting guidelines and ontology. Section 5 presents some suggested extensions for the TBX2RDF model and finally, in Section 6, we conclude and suggest future work.

2. Deutsche Bahn Terminology

The Deutsche Bahn (DB)[6] terminology aims, amongst others, at supporting the group-wide linguistic consistency, as a uniform use of language is gaining in importance with internationalization. The DB Language Portal helps with the translation of railway and company-specific words. Its underlying terminology covers 16 languages and can be accessed online[7]. The

[1] TBX stands for “TermBase eXchange”. See https://www.tbxinfo.net/ [accessed 2021-10-24], See also (Lommel *et al*., 2014) and (Melby, 2015).

[2] RDF stands for “Resource Description Framework”. See https://www.w3.org/TR/rdf-primer/.

[3] An updated version of those guidelines is available at https://github.com/bpmlod/report/blob/gh-pages/multilingual-terminologies/index.html [accessed 2021-10-24].

[4] https://www.w3.org/community/ontolex/

[5] http://lider-project.eu/lider-project.eu/index.html [accessed 2021-10-24].

[6] Deutsche Bahn AG is a large German railway company.

[7] www.deutschebahn.com/dblanguageportal [accessed 2021-10-24].

data we are working on is a TBX file export for Multiterm[8] friendly provided for this study by the Language Department of the Deutsche Bahn. In this export, values of the "type" attribute are written in German language (one our intention is to propose a way to support multilingualism for those values). Examples of this export, covering here only German and French terms, are given in the Annex.

3. Interactive Terminology for Europe (IATE)

IATE is the multilingual terminology database of the European Union, intended to support EU translators and as a means of standardisation of terminology across all institutions. It is available through an open access platform and constitutes the most important terminological reference for translators and language users in Europe. It collects terminological data from ten European Union's partners, including the European Parliament, the European Commission, the Court of Justice of the EU, the European Central Bank, and the Translation Centre for the Bodies of the European Union (CdT), to mention but a few.

The database puts together terminologies that were previously generated by the above-mentioned partners, such as Euroterms from the CdT and CuriaTerm from the Court of Justice. This means that this resource is also multidisciplinary, including terms in the financial, environmental, and labour domain, amongst others. Overall, IATE contains more than 8 million terms in the 24 EU languages plus Latin.

4. LIDER Guidelines and TBX Ontology

The aim of the past LIDER project was "to provide the basis for the creation of a Linguistic Linked Data cloud that can support content analytics tasks of unstructured multilingual cross-media content"[9], also including the mapping TBX to RDF. LIDER developed for this purpose a series of guidelines[10]

[8] MultiTerm is a termbase management software, provided by SDL. See https://docs.rws.com/785445/641133/sdl-multiterm-2019/welcome-to-------------sdl-multiterm-2019 [accessed 2021-10-249].

[9] http://lider-project.eu/lider-project.eu/index.html [accessed 2021-10-24].

[10] An updated version of those guidelines is available at https://github.com/bpmlod/report/blob/gh-pages/multilingual-terminologies/index.html [accessed 2021-10-24].

and a TBX ontology[11], in which TBX elements are converted into OWL[12] and associated with other RDF vocabularies, while the basic vocabulary chosen as the backbone of the conversion was the *lemon* model[13]. (Cimiano *et al*., 2015) and (McCrae *et al*., 2015) describe the TBX2RDF approach and the resulting resources. (di Buono *et al*., 2020) presents recent developments related to this TBX to RDF initiative, which consist in transforming and publishing terminologies as linked data, relying on a virtualization approach that is making use of containerization technologies.

While the LIDER TBX2RDF approach is representing the TBX terminological concepts as skos:Concept and the TIG/NTIG elements of TBX as ontolex:LexicalEntry, most of the other TBX elements are straightforwardly mapped onto RDF, meaning that they are encoding as URIs for representing a resource that can be associated with RDF predicates and objects. We note also that TBX2RDF does not represent the TBX langSet data as such, but instead is creating language specific lexicons in which all the data included in the original langSet element are encoded.

In our transformation work, we are investigating if all of the original TBX elements can be modelled also by reusing existing vocabularies, in order to take more advantage of the graph-based modelling facility that is supported by RDF. In doing so, we can introduce sub-class hierarchies and support the translation of all elements of TBX, beyond the sole translation of terms. Our model can still be mapped backward to a native TBX representation.

5. TBX2RDF Extensions

In our current work we make use of the most recent version of OntoLex-Lemon, which is effectively integrating the SKOS vocabulary[14] for expressing conceptual units. This was not the case for the former *lemon* vocabulary, which was used in the LIDER project. But the difference in the modelling is minimal, as we can now use properties defined in OntoLex-Lemon for linking the concepts to lexical entries, while the LIDER TBX2RDF converter

[11] https://github.com/cimiano/tbx2rdf/blob/master/ontology/tbx.owl [accessed 2021-10-24].

[12] OWL stands for "Web Ontology Language". See https://www.w3.org/TR/owl2-primer/ [accessed 2021-02-13].

[13] See (McCrae *et al*., 2012) for more details.

[14] SKOS stands for "Simple Knowledge Organization System". It is a W3C recommendation as a "common data model for knowledge organization systems such as thesauri, classification schemes, subject heading systems and taxonomies" (https://www.w3.org/TR/skos-reference/ [accessed 2021-10-24]).

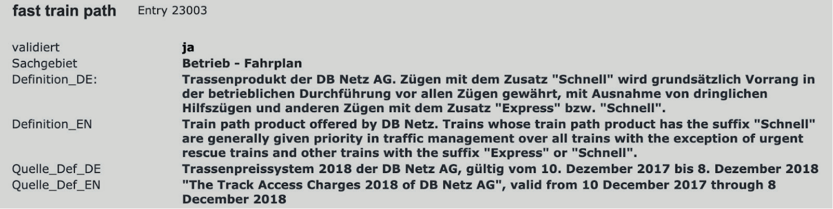
was using a custom property for this purpose. We introduce a skos:ConceptScheme for encoding the whole conceptual organisation of the DB terminology, and within this scheme we allow for the definition of specific domain subsets, something which is not explicitly foreseen in the original terminology, and maybe also not possible to formulate in TBX[15].

As the most recent version of OntoLex-Lemon is foreseeing a class ontolex-:LexicalConcept for linking lexical entries to the conceptual part described in the SKOS vocabulary, we encode all the term IDs as instances of this class, as this can be seen in FIG. 3.

Another, and more significant, departure from the LIDER TBX2RDF approach is the fact that we model definitions and contexts as instances of classes, and no longer as literal values, as was done previously by applying the property skos:definition. In doing so, we can describe specific relations between the definitions within one language or across different languages. In the latter case, we can specify if the definitions given for concepts in two different languages are translations of each other, are multilingual equivalents or are just monolingual definitions included in the multilingual terminology.

We have encountered those issues for the definitions and notes in both the DB and the IATE terminology resources. Within their entries, we can find data about term definitions and notes that cannot be represented only as literal values of properties. Modern approaches to generate terminologies from diverse resources also have an imperative need to represent the provenance of each piece of data, a feature that current vocabularies do not support.

FIG. 1 and FIG. 2 below are showing how definitions and notes are included in both the DB and the IATE terminologies.

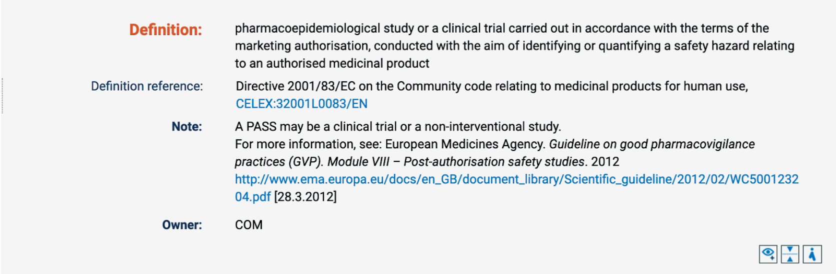


fast train path Entry 23003

validiert	**ja**
Sachgebiet	**Betrieb - Fahrplan**
Definition_DE:	**Trassenprodukt der DB Netz AG. Zügen mit dem Zusatz "Schnell" wird grundsätzlich Vorrang in der betrieblichen Durchführung vor allen Zügen gewährt, mit Ausnahme von dringlichen Hilfszügen und anderen Zügen mit dem Zusatz "Express" bzw. "Schnell".**
Definition_EN	**Train path product offered by DB Netz. Trains whose train path product has the suffix "Schnell" are generally given priority in traffic management over all trains with the exception of urgent rescue trains and other trains with the suffix "Express" or "Schnell".**
Quelle_Def_DE	**Trassenpreissystem 2018 der DB Netz AG, gültig vom 10. Dezember 2017 bis 8. Dezember 2018**
Quelle_Def_EN	**"The Track Access Charges 2018 of DB Netz AG", valid from 10 December 2017 through 8 December 2018**

FIG. 1 – *Term excerpt from the DB Portal showing definitions and their sources.*

[15] See (Reineke and Romary, 2019) for a discussion on the difference between the "subjectField" in TBX and the conceptual hierarchy in SKOS.



Definition: pharmacoepidemiological study or a clinical trial carried out in accordance with the terms of the marketing authorisation, conducted with the aim of identifying or quantifying a safety hazard relating to an authorised medicinal product

Definition reference: Directive 2001/83/EC on the Community code relating to medicinal products for human use, CELEX:32001L0083/EN

Note: A PASS may be a clinical trial or a non-interventional study.
For more information, see: European Medicines Agency. *Guideline on good pharmacovigilance practices (GVP). Module VIII – Post-authorisation safety studies*. 2012 http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/02/WC500123204.pdf [28.3.2012]

Owner: COM

FIG. 2 – *Entry from the IATE portal showing the definition of the term "post-authorization safety study" with a note with several elements.*

To overcome the above-mentioned gaps from the existing vocabularies, we have worked on a complementary module to model current terminological resources. We have published a first draft of this proposal as a wiki page within the Ontology Lexica Community Group. This extension proposal, temporarily named as Termlex, works with the ontolex:LexicalConcept as its key element. Therefore, one term ID is modelled as one LexicalConcept. This idea differs from the original intention of the Ontolex vocabulary where one LexicalConcept could have many LexicalSenses with several LexicalEntries associated. To avoid ambiguity, following traditional terminology theories, we restrict to one LexicalSense per LexicalConcept. Such a restriction can be represented with a functional property.

We present in this paper the aspects of Termlex dealing with definitions and notes, for which 3 classes are introduced:

Definition Class: a definition is a statement that explains the meaning of a concept. As mentioned earlier, we need the class Definition since this language data is much more than a literal. For instance, in FIG. 1, we see the terminology card of the term "fast train path", with two definitions and each of them has a different source that must be represented. Also, we propose a distinction amongst lexicographic and terminological definitions since their scope is also different. We therefore suggest creating the subclasses LexicographicDefinition and TerminologicalDefinition.

Note Class: the same happens with notes. A note may contain much more data than a simple text. In FIG. 2, we identify four different types of information in the note attached to the term "post-authorization safety study":

- The value (*A PASS may be a clinical trial…*), a literal that could be represented with the property rdf:value.
- A note for more information (*For more information, see…*), that could be modelled with a skos:note property, inside our Note class.

A pointer to the additional information (*http://ema.europa.eu…*), that could be modelled with a rdfs:seeAlso property.

The creation date of the note (*28.3.2012*) that could be represented with dcterms:created or simply dcterms:date.

Source Class: Likewise, sources can be much more than one piece of data: title, identifier, author, organisation, or application, to mention but a few. Therefore, we propose a class to collect them all. This class is especially needed when terminologies are generated from multiple resources, to maintain the traceability, since sources may even be chained, for instance, a definition that has been extracted with an application (source 1) from a given corpus (source 2). To model each element inside our Source class we also reuse other vocabularies such as DublinCore[16] or the Prov Ontology[17]. Our proposed inclusion of those classes with elements of OntoLex-Lemon is displayed in Fig. 3.

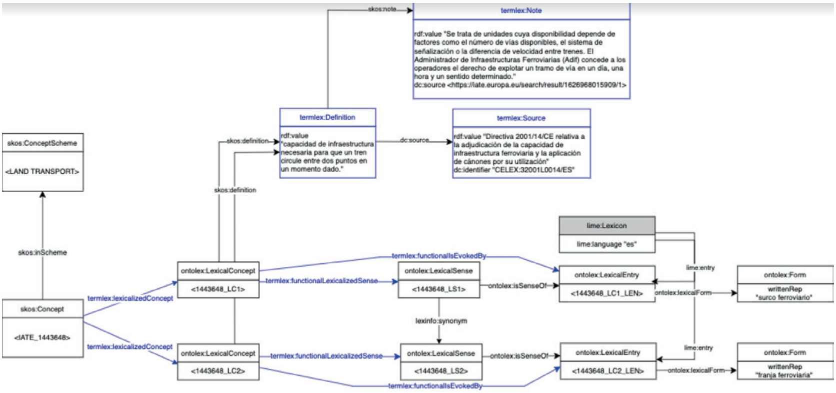

skos:note
termlex:Note
rdf:value "Se trata de unidades cuya disponibilidad depende de factores como el número de vías disponibles, el sistema de señalización o la diferencia de velocidad entre trenes. El Administrador de Infraestructuras Ferroviarias (Adif) concede a los operadores el derecho de explotar un tramo de vía en un día, una hora y un sentido determinado."
dc:source <https://iate.europa.eu/search/result/1626968015909/1>
termlex:Definition
rdf:value "capacidad de infraestructura necesaria para que un tren circule entre dos puntos en un momento dado."
dc:source
termlex:Source
rdf:value "Directiva 2001/14/CE relativa a la adjudicación de la capacidad de infraestructura ferroviaria y la aplicación de cánones por su utilización"
dc:identifier "CELEX:32001L0014/ES"
skos:definition
skos:definition
skos:ConceptScheme
<LAND TRANSPORT>
skos:inScheme
lime:Lexicon
lime:language "es"
termlex:functionalIsEvokedBy
ontolex:LexicalConcept
<1443648_LC1>
termlex:functionalLexicalizedSense
ontolex:LexicalSense
<1443648_LS1>
ontolex:isSenseOf
ontolex:LexicalEntry
<1443648_LC1_LEN>
lime:entry
ontolex:lexicalForm
ontolex:Form
writtenRep "surco ferroviario"
skos:Concept
<IATE_1443648>
termlex:lexicalizedConcept
termlex:lexicalizedConcept
lexinfo:synonym
ontolex:LexicalConcept
<1443648_LC2>
termlex:functionalLexicalizedSense
ontolex:LexicalSense
<1443648_LS2>
ontolex:isSenseOf
ontolex:LexicalEntry
<1443648_LC2_LEN>
lime:entry
ontolex:lexicalForm
ontolex:Form
writtenRep "franja ferroviaria"
termlex:functionalIsEvokedBy

Fig. 3 – *Integration of the classes Definition, Note and Source and the OntoLex-Lemon classes as proposed in Termlex*

[16] See for more details: https://dublincore.org/ [accessed 2021-10-24]. We use for example DublinCore for marking the creation date (dcterms:created or dcterms:date).

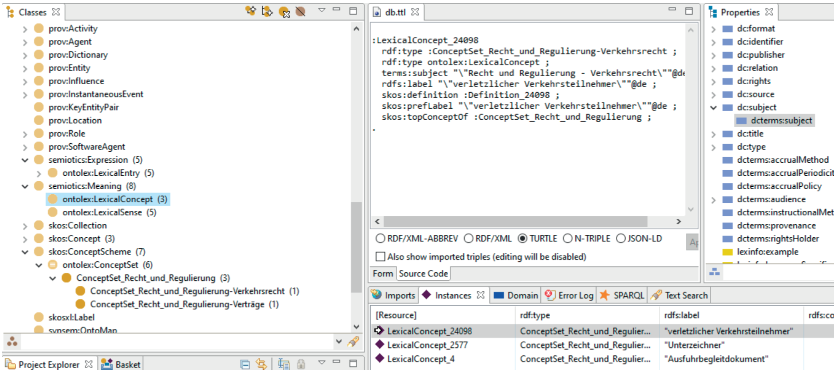
[17] See for more details: https://www.w3.org/TR/prov-o/ [accessed 2021-10-24].

We have also reviewed other vocabularies to reuse their classes to represent other elements of terminological entries, such as the term contexts, for which we are reusing the class lexicog:UsageExample, from the Lexicog[18] extension of Ontolex to model lexicographic resources; and the recommended usage of a term, for which we are reusing the class lexinfo:NormativeAuthorization, from the Lexinfo model[19], that has a fixed list of values (preferredTerm, deprecatedTerm, admittedTerm…).

Additionally, we introduce an explicit class hierarchy for the subdomains of the original DB terminology. There, the hierarchical relations between subject fields are represented by the sole use of a "hyphen" sign, like:

- terms:subject "\"Recht und Regulierung\""@de-DE
- terms:subject "\"Recht und Regulierung - Verkehrsrecht\""@de-DE
- terms:subject "\"Recht und Regulierung - Verträge\""@de-DE;

FIG. 4 displays a screenshot from the ontology editor we are using and shows, in the left panel, the subclass hierarchy we introduced, in our OWL-based representation of the original DB terminology. This way, we might apply inference mechanisms to all the subclasses of a specific domain of the terminology.



```
prov:Activity
prov:Agent
prov:Dictionary
prov:Entity
prov:Influence
prov:InstantaneousEvent
prov:KeyEntityPair
prov:Location
prov:Role
prov:SoftwareAgent
semiotics:Expression (5)
  ontolex:LexicalEntry (5)
semiotics:Meaning (8)
  ontolex:LexicalConcept (3)
  ontolex:LexicalSense (5)
skos:Collection
skos:Concept (3)
skos:ConceptScheme (7)
  ontolex:ConceptSet (6)
    ConceptSet_Recht_und_Regulierung (3)
      ConceptSet_Recht_und_Regulierung-Verkehrsrecht (1)
      ConceptSet_Recht_und_Regulierung-Verträge (1)
skosxl:Label
synsem:OntoMap
```

```
:LexicalConcept_24098
  rdf:type :ConceptSet_Recht_und_Regulierung-Verkehrsrecht ;
  rdf:type ontolex:LexicalConcept ;
  terms:subject "\"Recht und Regulierung - Verkehrsrecht\""@de
  rdfs:label "\"verletzlicher Verkehrsteilnehmer\""@de ;
  skos:definition :Definition_24098 ;
  skos:prefLabel "\"verletzlicher Verkehrsteilnehmer\""@de ;
  skos:topConceptOf :ConceptSet_Recht_und_Regulierung ;
.
```

○ RDF/XML-ABBREV ○ RDF/XML ● TURTLE ○ N-TRIPLE ○ JSON-LD

□ Also show imported triples (editing will be disabled)

```
dc:format
dc:identifier
dc:publisher
dc:relation
dc:rights
dc:source
dc:subject
  dcterms:subject
dc:title
dc:type
dcterms:accrualMethod
dcterms:accrualPeriodicit
dcterms:accrualPolicy
dcterms:audience
dcterms:instructionalMet
dcterms:provenance
dcterms:rightsHolder
lexinfo:example
```

[Resource]	rdf:type	rdfs:label	rdfs:co
LexicalConcept_24098	ConceptSet_Recht_und_Regulier...	"verletzlicher Verkehrsteilnehmer"	
LexicalConcept_2577	ConceptSet_Recht_und_Regulier...	"Unterzeichner"	
LexicalConcept_4	ConceptSet_Recht_und_Regulier...	"Ausfuhrbegleitdokument"	

FIG. 4 – *Representing hierarchies within domains of the DB terminology*

[18] See for more details: https://www.w3.org/2019/09/lexicog/

[19] See for more details: https://lexinfo.net/

6. Conclusions and Future Work

We described on-going work in porting the multilingual terminologies of the Deutsche Bahn and of IATE onto a Linked Data compliant representation language. This work led us to the question if it would not be suitable to extend the modelling of TBX terminologies in RDF already proposed by the LIDER TBX2RDF converter. One aspect consists in considering definitions or contexts as full ontological elements that can thus be put explicitly in relation to each other.

Another aspect we are working on is dealing with the inclusion of explicit subclass hierarchies, supporting the application of inference mechanisms to the specialisations of domains of the terminology.

Current work is dealing with the formulation of lexical restrictions that can apply to the terms, along the lines of the approach described in (Declerck *et al*., 2020).

Our proposal is now under discussion for admission as a new module extension to the OntoLex-Lemon framework.

Acknowledgements

Contributions by the German Research Center for Artificial Intelligence (DFKI GmbH) were supported in part by the H2020 project Prêt-à-LLOD with Grant Agreement number 825182. Contributions by the Austrian Centre for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences were supported in part by the H2020 project "ELEXIS" with Grant Agreement number 731015. The work described in this paper was also pursued in part in the larger context of the COST Action CA18209 - NexusLinguarum "European network for Web-centred linguistic data science".

References

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. (2016). Lexicon Model for Ontologies: Community Report. W3C Community Group Final Report

Maria Pia di Buono, Philipp Cimiano, Mohammad Fazleh Elahi, Frank Grimm. (2020). Terme-à-LLOD: Simplifying the Conversion and Hosting of Terminological Resources as Linked Data. In Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)

Cimiano, P., McCrae, J. P., Rodríguez-Doncel, V., Gornostay, T., Gómez-Pérez, A., Siemoneit, B., and Lagzdins, A. (2015). Linked terminologies: applying linked data principles to terminological resources. In Proceedings of the eLex 2015 Conference, pages 504-517.

John McCrae, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr and Tobias Wunner, *Language Resources and Evaluation*, 46(6), pp 701-709, (2012).

McCrae, J. P., Cimiano, P., and Doncel, V.R. (2015). Guidelines for linguistic linked data generation: Multilingual terminologies (TBX) -- Final Community Group Report. https://www.w3.org/2015/09/bpmlod-reports/multilingual-terminologies/

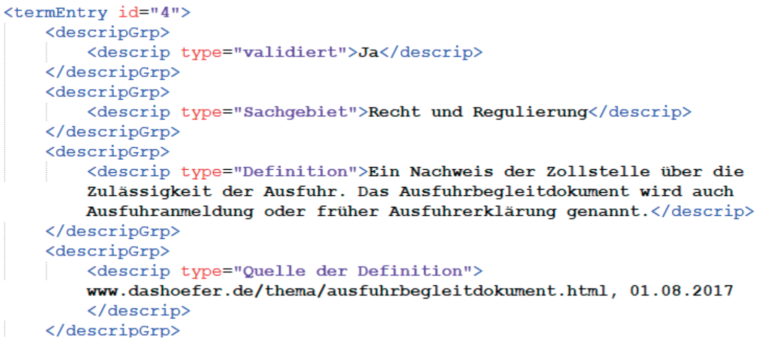
Declerck, Thierry, Racioppa, Stefania, Wissik, Tanja. 2020. Formal Representation of Linguistic Restrictions expressed in Terminologies. In Proceedings of TOTh 2020.

Lommel, Arle, Melby, Alan, Glenn, Nathan, Hayes, James and Snow, Tyler. 2014. TBX-Min: A Simplified TBX-Based Approach to Representing Bilingual Glossaries. Proceedings of *Terminology and Knowledge Engineering 2014*, Jun 2014, Berlin, Germany. 10 p.

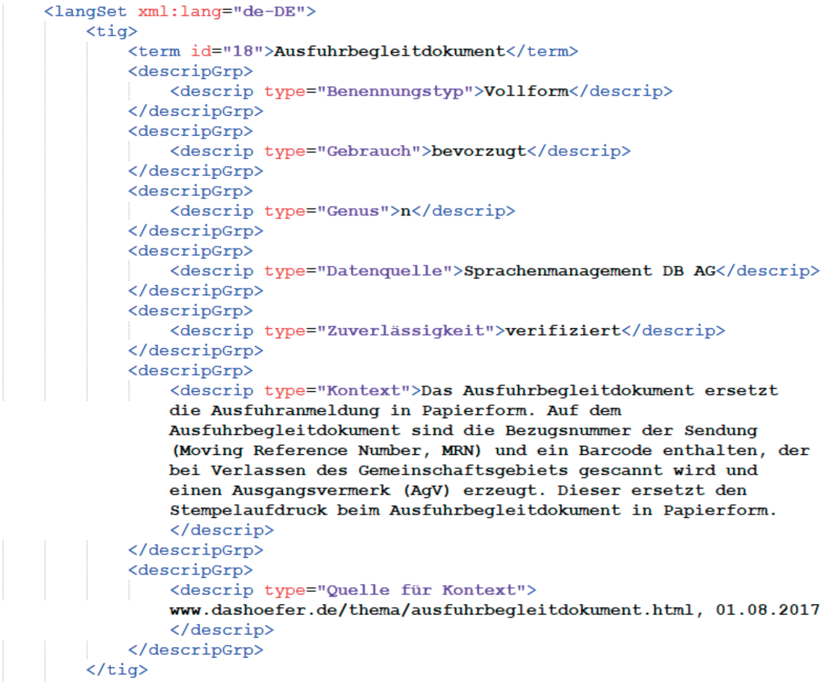
Melby, Alan K. 2015. TBX: A terminological exchange format for the translation and localisation industry. In Kockaert, Hendrik J. and Frieda Steurs (eds.). Handbook of Terminology. Volume 1. Amsterdam, Philadelphy: John Benjamins Publishing.

Detlef Reineke, Laurent Romary. Bridging the gap between SKOS and TBX. Die Fachzeitschrift für Terminologie, Deutscher Terminologie-Tag e.V. (DTT), 2019, Begriffssysteme und ihre Darstellung, 19 (2).

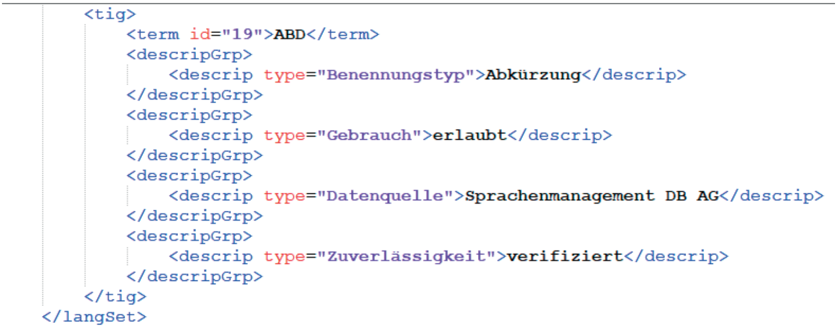
Annex: Examples from the TBX Export of the Deutsche Bahn Terminology (German and French)



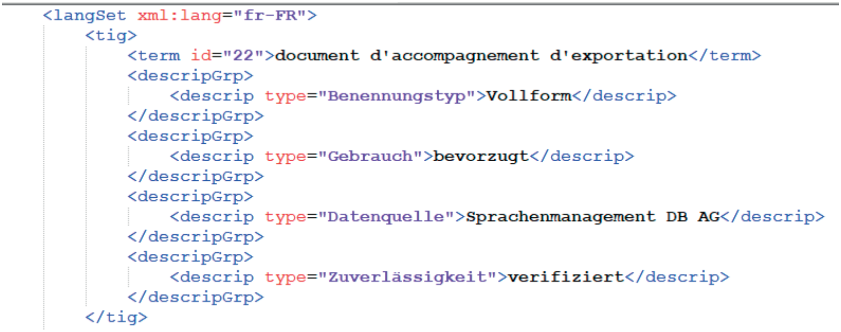
```
<termEntry id="4">
    <descripGrp>
        <descrip type="validiert">Ja</descrip>
    </descripGrp>
    <descripGrp>
        <descrip type="Sachgebiet">Recht und Regulierung</descrip>
    </descripGrp>
    <descripGrp>
        <descrip type="Definition">Ein Nachweis der Zollstelle über die
        Zulässigkeit der Ausfuhr. Das Ausfuhrbegleitdokument wird auch
        Ausfuhranmeldung oder früher Ausfuhrerklärung genannt.</descrip>
    </descripGrp>
    <descripGrp>
        <descrip type="Quelle der Definition">
        www.dashoefer.de/thema/ausfuhrbegleitdokument.html, 01.08.2017
        </descrip>
    </descripGrp>
```



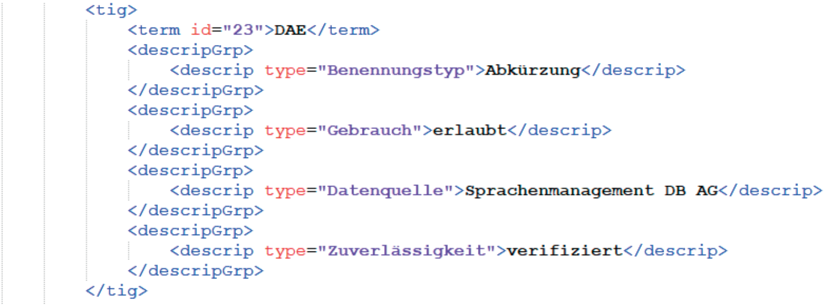
```
<langSet xml:lang="de-DE">
    <tig>
        <term id="18">Ausfuhrbegleitdokument</term>
        <descripGrp>
            <descrip type="Benennungstyp">Vollform</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Gebrauch">bevorzugt</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Genus">n</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Datenquelle">Sprachenmanagement DB AG</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Zuverlässigkeit">verifiziert</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Kontext">Das Ausfuhrbegleitdokument ersetzt
            die Ausfuhranmeldung in Papierform. Auf dem
            Ausfuhrbegleitdokument sind die Bezugsnummer der Sendung
            (Moving Reference Number, MRN) und ein Barcode enthalten, der
            bei Verlassen des Gemeinschaftsgebiets gescannt wird und
            einen Ausgangsvermerk (AgV) erzeugt. Dieser ersetzt den
            Stempelaufdruck beim Ausfuhrbegleitdokument in Papierform.
            </descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Quelle für Kontext">
            www.dashoefer.de/thema/ausfuhrbegleitdokument.html, 01.08.2017
            </descrip>
        </descripGrp>
    </tig>
```



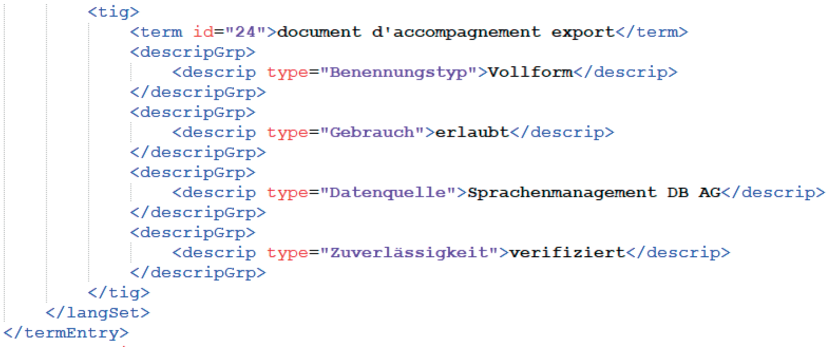
```
    <tig>
        <term id="19">ABD</term>
        <descripGrp>
            <descrip type="Benennungstyp">Abkürzung</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Gebrauch">erlaubt</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Datenquelle">Sprachenmanagement DB AG</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Zuverlässigkeit">verifiziert</descrip>
        </descripGrp>
    </tig>
</langSet>
```



```
<langSet xml:lang="fr-FR">
    <tig>
        <term id="22">document d'accompagnement d'exportation</term>
        <descripGrp>
            <descrip type="Benennungstyp">Vollform</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Gebrauch">bevorzugt</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Datenquelle">Sprachenmanagement DB AG</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Zuverlässigkeit">verifiziert</descrip>
        </descripGrp>
    </tig>
```



```
    <tig>
        <term id="23">DAE</term>
        <descripGrp>
            <descrip type="Benennungstyp">Abkürzung</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Gebrauch">erlaubt</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Datenquelle">Sprachenmanagement DB AG</descrip>
        </descripGrp>
        <descripGrp>
            <descrip type="Zuverlässigkeit">verifiziert</descrip>
        </descripGrp>
    </tig>
```



```
        <tig>
            <term id="24">document d'accompagnement export</term>
            <descripGrp>
                <descrip type="Benennungstyp">Vollform</descrip>
            </descripGrp>
            <descripGrp>
                <descrip type="Gebrauch">erlaubt</descrip>
            </descripGrp>
            <descripGrp>
                <descrip type="Datenquelle">Sprachenmanagement DB AG</descrip>
            </descripGrp>
            <descripGrp>
                <descrip type="Zuverlässigkeit">verifiziert</descrip>
            </descripGrp>
        </tig>
    </langSet>
</termEntry>
```

Transformation of Graphical Semiotic Models into a Graph-based Formal Representation

Thierry Declerck*/**, Marija Ivanović***

*DFKI GmbH, Saarland Informatics Campus D3 2
Multilinguality and Language Technology Lab
Stuhlsatzenhausweg, 3
D-66123 Saarbrücken, Germany
declerck@dfki.de
https://www.dfki.de/~declerck /
**Austrian Academy of Sciences
Austrian Centre for Digital Humanities and Cultural Heritage
Sonnenfelsgasse, 19
A-1010 Vienna, Austria
declerck@dfki.de
https://www.oeaw.ac.at/acdh/team/
***University of Vienna
Centre for Translation Studies
Gymnasiumstraße 50
A-1190 Vienna, Austria
marija.ivanovic@univie.ac.at

Abstract. Graphical models are often proposed for the purpose of explanative visualisation in different theories of signs, meaning, concepts, or references. We describe work dealing with the transformation of such graphical models into a graph-based formal representation language. The final aim of this formal representation is to be able to mark elements of those different theories as being compatible, complementary, or disjunct.

1. Introduction

The meaning (or semantic, or semiotic) triangle proposed by Ogden and Richards (1923) is a widely known graphical model summarizing the views of the authors on how a “symbol” is related to a “referent” via a “thought or reference”[1]. There are related designations for those elements, like (in the same order) “sign”, “object”, “concept”, etc. The “idea” of using a triangle for such a purpose was not entirely new and has been exploited for example by Charles Sanders Peirce for representing relations between a “representamen” (or “sign”) and an “object” via an “interpretant”[2].

In this paper we discuss the use of an integrated graph-based formal representation for the encoding of different graphical semiotic models used for visualizing theories of the sign, meaning or references. Besides the triangle model we consider the so-called dyadic model proposed by de Saussure (1916), and the semiotic quadrangle introduced by Eugen Wüster (1959)[3] or the semantic quadrangle by G. P. Melnikov (as displayed by Wang (2016)). We also deal, tentatively, with proposals that are combining triangles, like those described in (Sowa, 2010) or (Roche, 2007).

The deployment of geometrical figures for visualising sign or meaning theories are directly “inviting” to the use of graph-based formalisms for their formal encoding. We can even consider the dyadic model of de Saussure as consisting of two nodes related by an edge. For the triangles and quadrangles, we can see the angles as nodes, and the names carried by those angles as the labels of the nodes. And the lines between the angles can be considered as (possibly labelled) edges between nodes. We are therefore opting for investigating the use of the Resource Description Framework (RDF)[4] and vocabularies based on it for offering a formal representation of such graphical semiotic models, which were conceived for describing the sign or the meaning. In the longer term, our work is dealing with the linking of formal representation models for lexicographic, terminological, and general knowledge data.

An inspiration and motivation for our work was given by “YAT” (“Yet Another Triangle”) which at a metalevel is describing the possible relations

[1] This graphical model, as this is the case for all other graphical models, is displayed in the Annex, here with the number Annex:2.

[2] We take this information from (Sowa, 2010), but we are not aware of a concrete triangle designed by Peirce himself. (Chandler, 2007, page 30) is proposing such a concrete triangle, as suggested by one of his students. This graphical representation is given in Annex:3.

[3] (Ivanović, 2020) gives a detailed description of the various quadrangles introduced by Wüster. We note that the original German text for “semiotic” quadrangle is “vierteiliges Wortmodell” (*four-part word model*), where the “word” is thus playing a central role.

[4] For RDF, see https://www.w3.org/TR/rdf-primer/. For RDF-based vocabularies, see (among others) https://www.dnb.de/EN/Professionell/Metadatendienste/Exportformate/RDF-Vokabulare/rdf.html.

between 3 types of languages, as described by Roche (2007). We display in FIG. 1 this "meta" triangle.

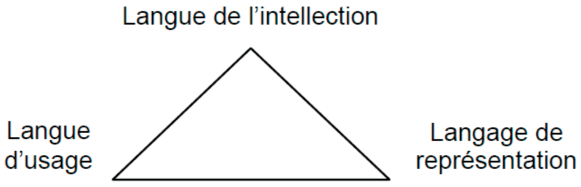


FIG. 1 – *The "meta" triangle by Roche (2007)*

We situate our work in the realm of the "Langage de représentation", but with the aim to describe relations that are existing between those three types of languages. We note that in this triangle the word "langue" is used for labelling two angles, and "langage" for one angle. The French language allows this distinction, which in this specific usage by Roche is marking the abstract meaning of a language as a (formalized) encoding system ("langage"), while the "langue" is the language as used in communicative situations. In fact, as can be seen in the next sections, we use a specific "langage de représentation" for encoding elements of both the "langue de l'intellection" and the "langue d'usage", relating them to conceptual and ontological elements that are encoded in the same type of "language de représentation".

2. RDF, SKOS, and OntoLex-Lemon

The Resource Description Framework (RDF) is a W3C recommendation that was developed for easing the exchange of data, with a focus on interoperability. The basic structure of RDF is a triple expressing a subject and its relation to an object, representing in fact a graph-based model. On the top of RDF, more complex and expressive knowledge representation languages have been designed, also in the context of the so-called W3C Semantic Web Stack[5].

[5] See https://en.wikipedia.org/wiki/Semantic_Web_Stack for an introduction to the Semantic Web Stack.

RDF(s)[6] and OWL[7] are supporting a higher expressivity and reasoning in the representation language, while remaining in the paradigm of the triple as the basic modelling approach.

SKOS (Simple Knowledge Organization System) is another W3C recommendation, yet an RDF-based vocabulary that was developed as “a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary.”[8]

OntoLex-Lemon (Cimiano *et al.* 2016) is another RDF-based vocabulary which was originally designed for modelling natural language expressions used in the labels of OWL, RDF(s) and RDF encoded ontologies. But extensions to the basic OntoLex-Lemon model made it also suitable for lexicographic purposes[9], and the inclusion of the SKOS vocabulary put OntoLex-Lemon in the position of internally linking lexical elements with conceptual elements included in terminologies, thesauri, and other conceptual schemes. The picture in Fig. 2 gives an overview of the core module of OntoLex-Lemon.

In the following sections, we show how this model can be used for representing the different types of graphical models of the sign or meaning theories we have mentioned in Section 1. For this purpose, we will in each case use the labels displayed in Fig. 2.

[6] RDF(s), or also RDFS, RDF-S, RDF/S, stands for RDF Schema. See https://www.w3.org/TR/rdf-schema/ for more details.

[7] OWL stands for Web Ontology Language. See https://www.w3.org/TR/owl2-primer/.

[8] See https://www.w3.org/TR/skos-primer/ for more details.

[9] See the OntoLex-Lemon “lexicog” module: https://www.w3.org/2019/09/lexicog/ and (Bosque Gil *et al.*, 2019).

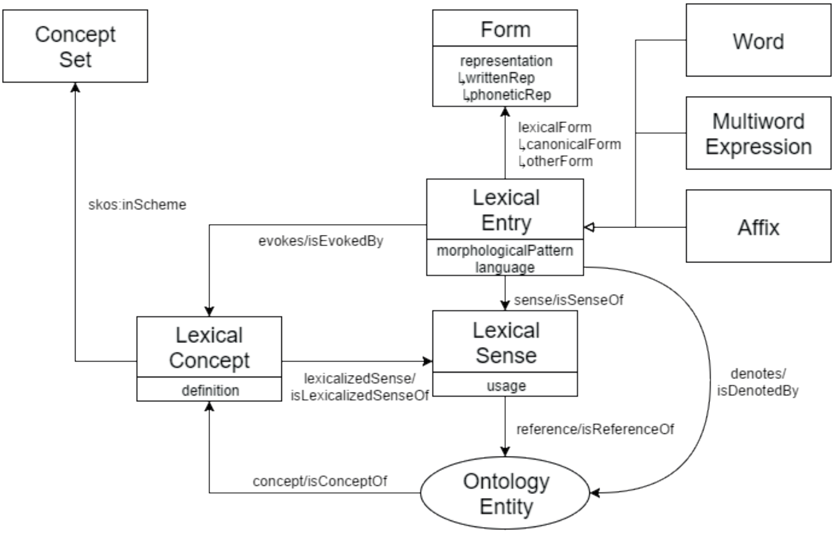


Fig. 2 – *The core module of OntoLex-Lemon, taken from https://www.w3.org/2016/05/ontolex/.*

3. Modelling the graphical Models of the Sign Theories in OntoLex-Lemon

In the following sections we concentrate on four upper classes of OntoLex-Lemon, not taking into account the ontolex:ConceptSet class and the ontolex:Form class. This results in a kind of a "quadrangle", which is displayed in Fig. 3 below. We will "decompose" this quadrangle in all possible triangles included in it to represent the different graphical semantic/semiotic models we can deal with.

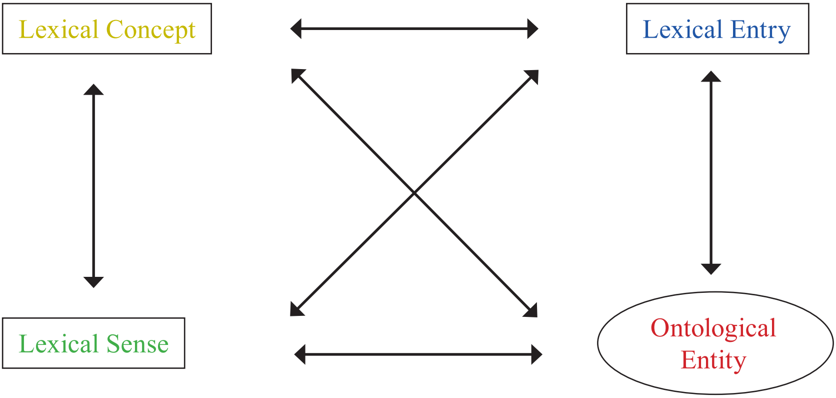


FIG. 3 – *Four upper classes of OntoLex-Lemon and the relations between them*

3.1. The Dyadic Models

We start with the dyadic model by de Saussure (see Annex:1). As quoted by Sowa (2010), this model places unity of the sign uniquely in the relation between a "sound-image" (which in OntoLex-Lemon would be integrated as a subclass of a Form related to a LexicalEntry) and a "concept", which in OntoLex-Lemon we encode as a LexicalConcept (a subclass of a skos:Concept). Sowa (2010) is contrasting this approach to model-theoretic semantics approaches which link every natural language expression to a real object in the world[10]. This different dyadic model can be expressed in OntoLex-Lemonby the property "denotes" which relates a lexical entry directly with an ontology entity.

But in doing so, OntoLex-Lemon still opens the possibility to establish links of the dyadic constituents to other elements of its model.

[10] Quoting Sowa (2010): "Tarski, Quine, and many other logicians […] focused on the dyadic link between the sign and object."

3.2. The Triadic Model

The semantic triangle by Ogden and Richards (1923), displayed in Annex:2, contains interestingly a disjunction in the label associated with its top angle: “THOUGHT OR REFERENCE”. We suggest therefore at this place two OntoLex-Lemon “triangles” for possibly representing the triadic model by Ogden and Richards. One mediating the symbol (LexicalEntry) with the object (Ontology) via the “thought” (LexicalConcept) and one via the “reference” (LexicalSense), as this is visualized in Fig. 4. While the triangle by Ogden and Richards is not foreseeing a direct relation between the symbol and the object, we have the possibility in OntoLex-Lemon to link a LexicalEntry directly to an Ontology Entity, via the ontolex:denotes property.

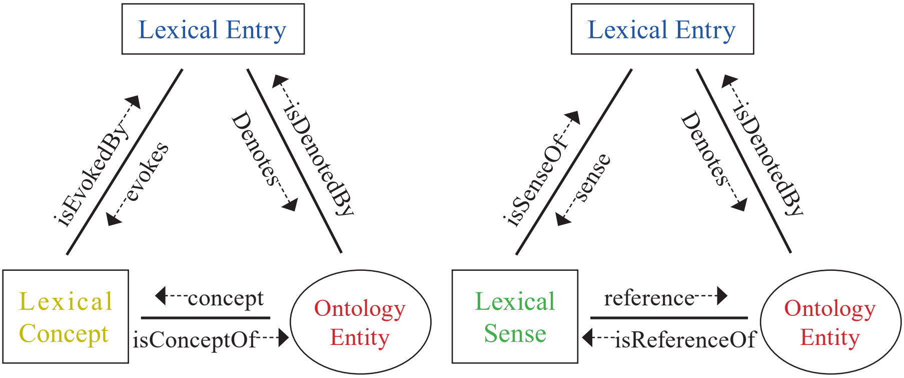


Fig. 4 – *Two “ontolex” triangles for representing the Ogden-Richards semiotic triangle, taking into account the disjunction “THOUGHT OR REFERENCE” in its top angle. We represent this disjunction with the classes “ontoloex:LexicalConcept” (left) and “ontolex:LexicalSense” (right).*

At this stage of our work, we do not have yet a proposal for representing the “Peircean triangle” (Annex:3) in OntoLex-Lemon and related RDF vocabularies.

3.3. The Semiotic/Semantic Quadrangles

In the case of the semantic quadrangle by G. P. Melnikov (Annex:4. a.), we again note the dotted lines between the symbol (sign) and the object (thing).

But in OntoLex-Lemon, we do have the "ontolex:denotes" property that allows to link directly an ontolex:LexicalEntry to an ontology entity (standing for real world objects) as a direct relation, which is parallel to the indirect relation mediated by an ontolex:LexicalSense. We refer here to FIG. *2* for the OntoLex-Lemon visualisation.

We display in below the other 2 triangles that can be extracted from the quadrangle displayed in FIG. *2*, so that all the possible "ontolex" triangles (which are also included in the semantic quadrangle) are rendered in this paper, showing that all the aspects of this quadrangle can be represented in OntoLex-Lemon, even those cases in which neither a lexical entry nor an ontological entity are involved.

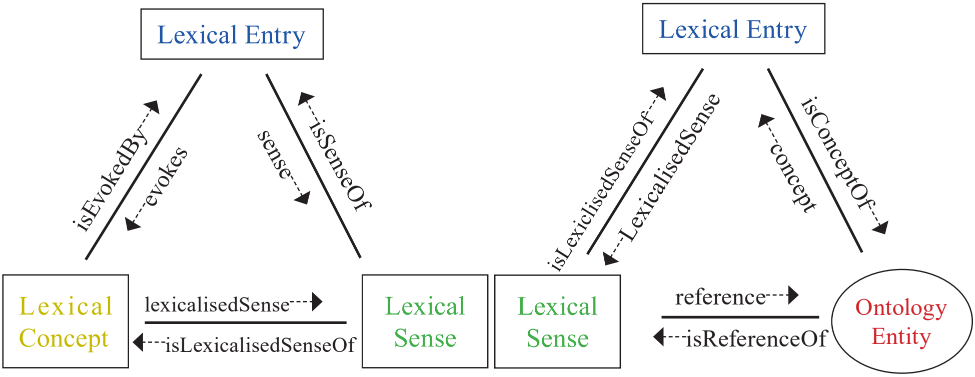


FIG. 5 – *The remaining two triangles that can be extracted from the OntoLex-Lemon "quadrangle", one without considering the real world (ontology entity) and one not considering the symbol (lexical entry)*

We are currently working on the transposition of the various quadrangles by Wüster into OntoLex-Lemon.

3.4. The Combined Triangles

Some authors propose a combination of triangles, like Roche (2007 – Annex:5)[11] and Sowa (2010 – Annex:6). Those proposals are highly interest-

[11] This graphical model is a corrected version of another graphical model presented in the same publication, and which is displayed in Annex:7.

ing as they allow to formulate relations between triangles, visualizing additional aspects of the sign, the meaning, or the references. While Roche (2007) rather establishes a conceptual relation between two triangles, Sowa (2010) introduces a kind of sequential relations between triangles, which can be horizontal (marking possibly a temporal sequence) or vertical. We deal in this paper only with the proposal by Roche (2007).

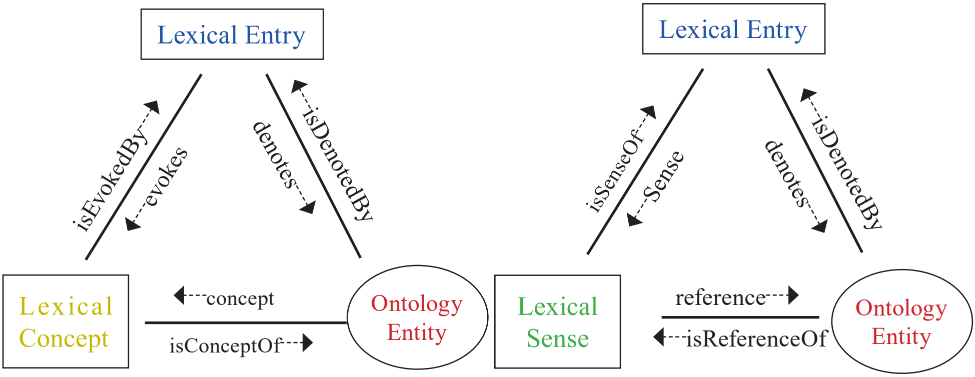


Fig. 6 –??

In our opinion, the two "OntoLex-Lemon triangles" displayed Fig. 4 (and repeated just above) are very close to the latest version the double triangle by Roche (2007), which is shown in Annex:5. But we need to get into more details to explain our view. It is also useful to study a former version by Roche of this double triangle, which is displayed in Aneex:7. There we can see that the relation between the "signifiant" (in the left triangle) and the "identifiant" (in the right triangle) is represented as a set intersection. This type of representation disappeared in the double triangle displayed in Annex:5 (where we notice that both triangles have exchanged their position, but we see just a practical reason for this exchange of positions). Common to both versions of the double triangle is that the triangles are related to each other by a set intersection representation. We can assume that this set intersection is the type of relations that is existing between "concept" and "signifié" on the one hand, and "identifiant" and "signifiant" on the other hand. Those relations are represented by double arrows in the graphic displayed in Annex:5.

We suggest in a first step to specialise the inverse OntoLex-Lemon property "lexicalizedSense/isLexicalizedSenseOf", which is linking a LexicalConcept to a LexicalSense, to a functional property. This can ensure the uniqueness of interpretation out of a multitude of senses possibly associated with lexical items used in the context of a terminology, which requires an unambiguous lexical realisation of its concepts. This would respond to the top double arrow in the double triangle by C. Roche. Concerning the bottom double arrow in this double triangle, one solution could be to extend the OntoLex/Lemon, adding a "terminological entry", as a parallel element to the "lexical entry", as suggested in (Kudashev and Kudasheva, 2010), who are using the expressions "terminological lexeme" and "lexicographic lexeme". But our preference goes in not duplicating elements, and we think that the dense interlinking of a LexicalEntry (and all its forms) with a Lexical Concept and a LexicalSense, whereas the linking between a lexical concept and a lexical sense would be restricted to have maximally one instance, should suffice to distinguish a "terminological" entry from a "lexicographic" one. This way, we can deal with the distinction operated by C. Roche between an "identifiant" and a "signifiant", while the "identifiant" can remain encoded as non-lexical code in our RDF-based representation.

4. Conclusions and Future Work

We presented on-going work in re-using the OntoLex-Lemon model and associated RDF vocabularies for formally representing a series of graphical visualizations of different theories of the sign, the meaning, or the references, which are playing a role in the way terminology is situating itself in comparison and cooperation with other fields that have the language at their core, as this is exemplified in the work by C. Roche (2007) on establishing an "onto-terminology".

While there are still some instances of graphical models to be dealt with, our future efforts will also go in investigating the concrete impact of our work, as it could be offering a meta description of the relation between computational ontologies, terminologies and lexicons, a programme already described in (Roche, 2007), where a graphical representation of the relation between the languages of representation, of usage and of thought is given.

Acknowledgements. Contributions by the German Research Center for Artificial Intelligence (DFKI GmbH) were supported in part by the H2020 project Prêt-à-LLOD with Grant Agreement number 825182. Contributions by the Austrian Centre for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences were supported in part by the H2020 project

“ELEXIS” with Grant Agreement number 731015. The work described in this paper was also pursued in part in the larger context of the COST Action CA18209 - NexusLinguarum “European network for Web-centred linguistic data science”.

References

Bosque Gil, J., Lonke, D., Gracia del Río, J., Kernerman, I. (2019). Validating the OntoLex-lemon Lexicography Module with K Dictionaries' Multilingual Data. In Proceedings of eLex 2019.

Chandler, D. (2007, second edition). Semiotics: the basics. Routledge.

de Saussure, F. (1916, 1971). *Cours de linguistique générale*. Payot, Paris, 1971. Exporté de Wikisource le 05/13/20. [accessed online: https://fr.wikisource.org/wiki/Cours_de_linguistique_g%C3%A9n%C3%A9rale/Texte_entier]

Ivanović, M. (2020). Eugen Wüster's Semiotic Quadrangles. Proceedings of the TOTh 2020 conference.

Kudashev, I., Kudasheva, I. (2010). Semiotic Triangle Revisited for the Purposes of Ontology-based Terminology Management. In proceedings of the TOTh onférence TOTh 2010

Ogden, C. K., and Richards, I. A. (1923, 1989). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*, with supplementary essays by Bronislaw Malinowski and F. G. Crookshank, Introduction by Umberto Eco. San Diego: Harcourt Brace Jovanovich. 1st ed., 1923.

Roche, C. (2007). Le terme et le concept: fondements d'une ontoterminologie, In Conférence TOTh 2007 -Terminologie & Ontologies: Théories et Applications. Annecy, Institut Porphyre

Sowa, J.F. (2010). The Role of Logic and Ontology in Language and Reasoning. In Theory and Applications of Ontology: Philosophical Perspectives; Poli, R., Seibt, J., Eds.; Springer: Dordrecht, The Netherlands, 2010; pp. 231-263; ISBN 978-90-481-8845-1. (we quote here from a slightly modified version, available at http://www.jfsowa.com/pubs/rolelog.pdf [accessed 2021-02-199]).

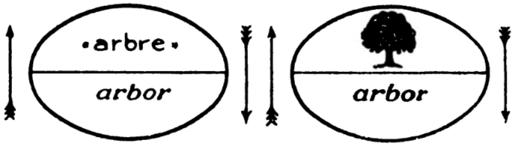
Wang. M. (2016). Toward the Meaning of Linguistic Signs: A Hierarchical Theory. Language and Semiotic Studies, Vol 2, N. 1.

Wüster, E. (1959, 2001). Das Worten der Welt, schaubildlich und terminologisch dargestellt. In *Terminologie und Wissensordnung. Schriften aus dem*

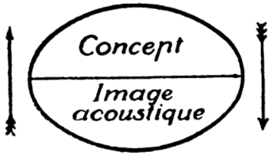
Gesamwerk von Eugen Wüster, edited by Heribert Picht and Klaus-Dirk Schmitz, 21-52. Wien: TermNet.

Annex: Dyadic Models, Triangles, Quadrangles and Compositions

1. Dyadic model of de Saussure

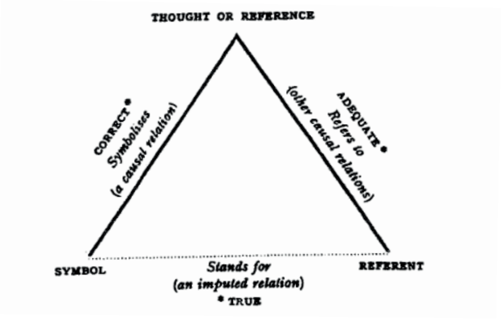


https://fr.wikisource.org/wiki/Cours_de_linguistique_g%C3%A9n%C3%A9rale/Texte_entier#/media/Fichier:Saussure-cours-p-099b.png



https://fr.wikisource.org/wiki/Cours_de_linguistique_g%C3%A9n%C3%A9rale/Texte_entier#/media/Fichier:Saussure-cours-p-099a.png

2. Triadic Model of Ogden and Richards



https://courses.media.mit.edu/2004spring/mas966/Ogden%20Richards%201923.pdf, page 11

3. (possible) Triadic Model of Peirce

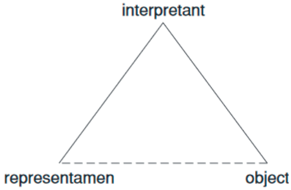
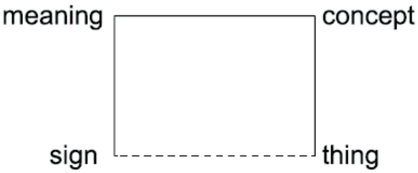


FIGURE 1.5 Peirce's semiotic triangle

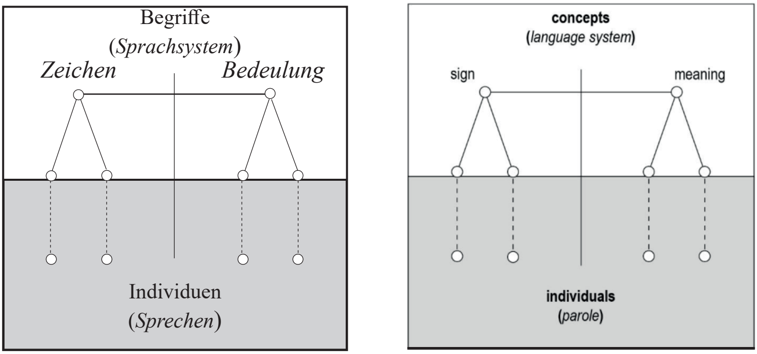
(Chandler, page 30, Image taken from http://www.wayanswardhani.lecture.ub.ac.id/files/2013/09/Semiotics-the-Basics.pdf)

4. Quadrangles

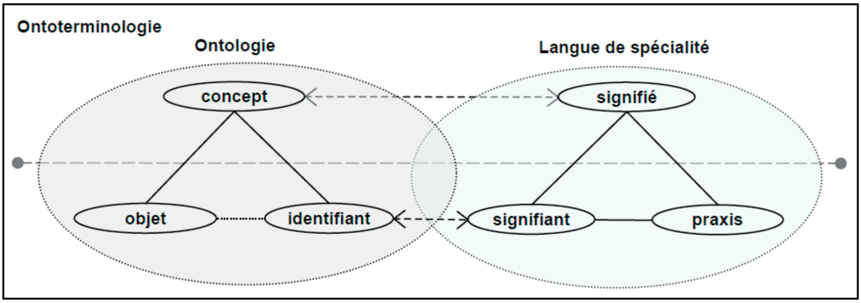
a. Sematic Quadrangle by G. P. Melnikov (taken from (Wang, 2016)



b. Semantic Quadrangle of Wüster (the original, on the left, as published in Wüster, 1959:308). The translation, on the right, taken from (Ivanović, 2020).

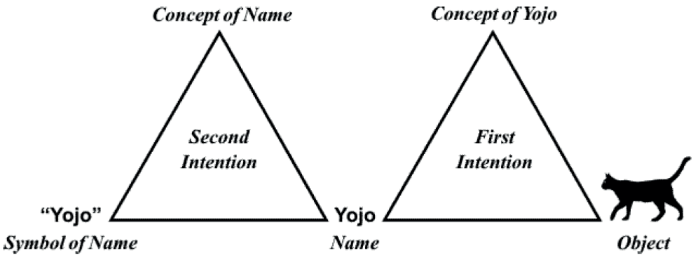


5. The double semantic triangle, by C. Roche (2007)



6. Combination of triangles, by J. Sowa (2010)

a. The “Scholastic Triangles” (Sowa, 2010: Figure 1)



b. (Peirce) concept of representation by means of two meaning triangles (Sowa, 2010: Figure 3)

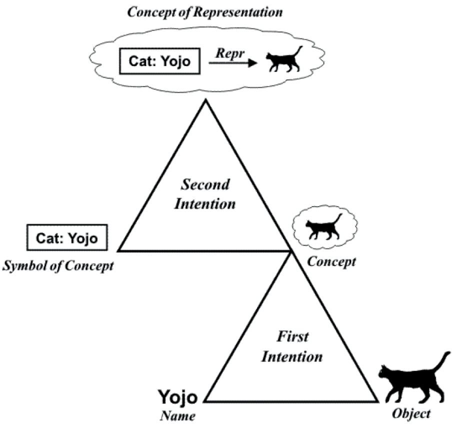
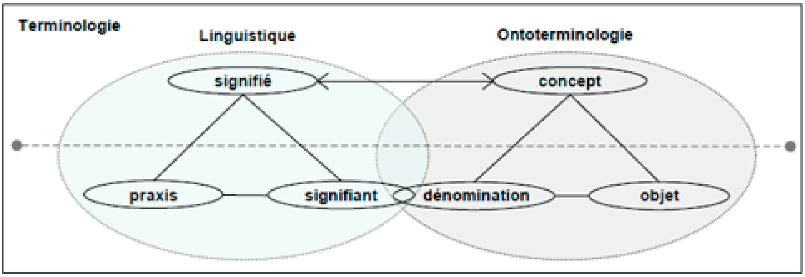


Figure 3. Meaning triangles for the concept of representation

7. The first version of the double semantic triangle, by C. Roche (2007)



Résumé

Des modèles graphiques sont souvent proposés à des fins de visualisation explicative pour différentes théories des signes, de la signification, des concepts ou des références. Nous décrivons notre travail portant sur la transformation de tels modèles graphiques en un langage de représentation formelle basé sur les graphes. L'objectif final de cette représentation formelle est de pouvoir marquer les éléments de ces différentes théories comme étant compatibles, complémentaires ou disjoints.

Role of the Corpus in Ontoterminology: the Case of the Balance of Payments

Stéphane Carsenty*, **

* LISTIC, Équipe Condillac, University Savoie Mont Blanc, France
stephane.carsenty@univ-smb.fr
** Language Services, Swiss National Bank, Switzerland

Abstract. The ontoterminological modelling of a domain uses a clearly conceptual approach to terminology, where the role of experts as a source for the elaboration of the system of concepts and for the validation of terms is essential. Nevertheless, ontoterminology can benefit from the use of corpora. This paper presents the issues linked with the construction of a text corpus in a socially constructed domain whose objects are immaterial, namely the balance of payments, and determines the role this corpus can play in the elaboration of the ontoterminology of this domain.

1. Introduction

Ontoterminological modelling, understood as the elaboration of a terminology whose concept system is a formal ontology, uses a clearly conceptual approach to terminology. The role of experts as a source for modelling the system of concepts and validating terms is essential. Nevertheless, ontoterminology can benefit from the use of corpora.

This paper aims at determining the role corpora can play in ontoterminology. We first shortly recall the basics of ontoterminology (section 2). Then we briefly describe the main features of the domain of the balance of payments, which is the domain on which this research focuses (section 3). After that, we present the issues linked with the current construction of a text corpus in that domain and metrics about it (section 4). Last, we analyse the role this corpus can play in the elaboration of the ontoterminology of this domain (section 5). In a short conclusion, we sum up the knowledge gained from these reflections and outline perspectives for future work.

2. Ontoterminology

Ontoterminology (Roche 2007a) is a paradigm introduced to take account of the increasing need for operationalisation of terminologies for IT applications in the digital society. Its purpose is to go beyond the mere goal of a "*clarification and standardisation of concepts and terminology for communication* *between humans*" (ISO 704:2009, Overview) and to permit "*a computational representation of concepts in terminology*" (Roche 2012) in order for terminologies to become interoperable.

An ontoterminology is a terminology whose concept system is a formal ontology (Roche 2012). "Formal" means that concepts are defined in a computer-readable language. They are built through specific differentiation: each concept is defined as a species of its superordinate concept, e.g. possessing all characteristics of the latter and an additional one, e.g. a delimiting characteristic that distinguishes it from its superordinate concept (ISO 1087-2019, 3.2.5). The former is thus linked to the latter with a generic relation (*is_a*). Other, non-generic relations may also be defined to enrich the concept system. Ontoterminology allows for the operationalisation of terminologies that ought to be consensual, consistent, precise, shareable, reusable and computable (Roche 2007a).

To model the ontoterminology of a domain, the terminologist groups essential characteristics that are stable enough to be designated with linguistic means (Roche & Papadopoulou 2020:60). Each set of such essential characteristics is a concept and is thus denoted by a lexical unit called a term. To be able to elaborate the (essential and non-essential) characteristics of concepts and the relations between concepts, the terminologist relies on domain knowledge. This is why experts are paramount.

Additionally, the two dimensions of terminology, namely the conceptual dimension and the linguistic one, are modelled distinctly because they are based on two different semiotic systems (Roche 2012). However, in order for terminology to be possible, "*these two systems must be both separated [...] and linked*". This is done, for this research, in a dedicated software environment, namely Tedi (http://ontoterminology.com/tedi).

In such a clearly conceptual approach where experts are central, we want to analyse the role corpora can play in a specific domain of application, namely the balance of payments.

3. Balance of payments

The balance of payments (BOP) belongs to the international accounts. It is a statistical statement published at regular intervals (monthly, quarterly and yearly) about all relationships of an economy[1] with the rest of the world (IMF 2009: 2.12, 9). Statisticians working at central banks or at statistics offices in each member state of the International Monetary Fund (IMF) establish the balance of payment. It consists of a systematic presentation of statistical aggregates recorded in international accounts referring to macroeconomic objects (like imports, exports, assets and liabilities, financial flows and positions, etc.). We have to do with immaterial objects (Carsenty 2020:84).

The domain of the BOP can be schematically represented with three semiotic triangles (see FIG 1): one at the bottom, which corresponds to the macroeconomic objects observed, one in the middle, which stands for the statistical aggregates computed by statisticians to represent the latter, and one at the top, which refers to the accounts in which the aggregates are recorded.

Like national accounts, government finance statistics and monetary and financial statistics, the balance of payments belongs to official statistics. It is a socially constructed domain (Carsenty 2020:85). To establish the BOP, statisticians follow rules set out by international organisations in dedicated reference documents and manuals. The International Monetary Fund defines the conceptual framework (IMF 2009) and the principles for implementing it (IMF 2014). These are based on and complemented by work done by the "*global official statistics community*" (UNECE 2019:3). This community includes the Statistical Office of the European Communities (Eurostat), the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), the United Nations Statistics Division and regional commissions of the United Nations, Secretariat and the World Bank (EC et. al. 2009:xlvii), the United Nations Economic Commission for Europe (UNECE), and national bodies. National statistics offices and central banks may adopt specific rules and principles in addition to the ones set out at international level, for example in order to comply with national or regional regulations. There is for example a European framework for establishing the BOP (e.g. OJEU 2012 and OJEU 2016).

[1] An economy is a national or a supranational entity with harmonised rules (e.g. a country, an economic union, like the European Union, or a currency union, like the euro area).

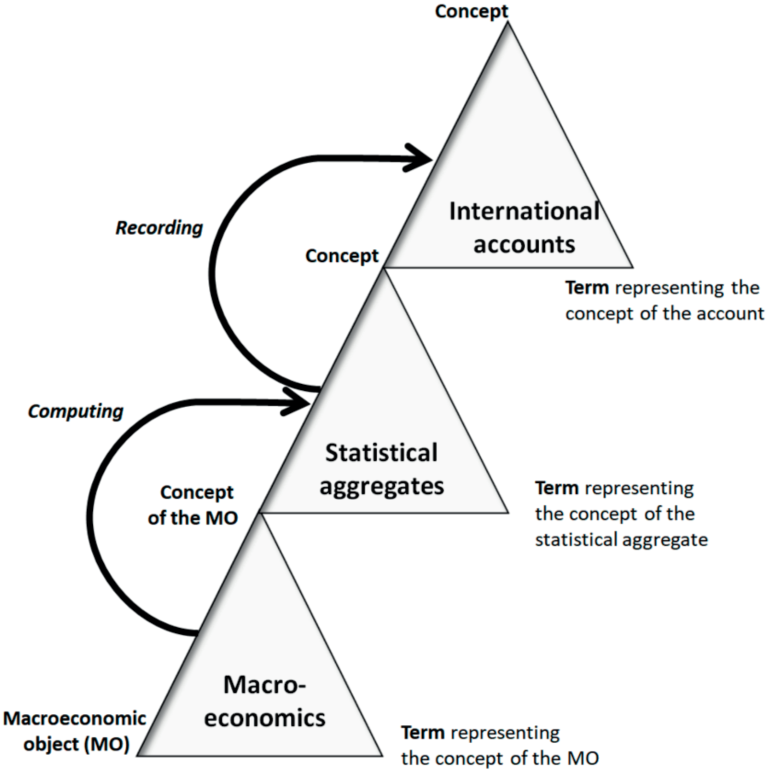

Concept
Recording
International accounts
Concept
Term representing the concept of the account
Computing
Statistical aggregates
Concept of the MO
Term representing the concept of the statistical aggregate
Macro-economics
Macroeconomic object (MO)
Term representing the concept of the MO

Fig. 1 – *Representation of the BOP after Sowa 2000:193 ff.*

International accounts, to which the BOP belongs, are based on the same theoretical principles as national accounts. Both are aimed at observing and analysing interactions between agents, and at aggregating them in institutional units grouped in institutional sectors, according to their connection to a specific economy (EC. *et al*. 2009:61 ff. and IMF 2009:50 ff.). Whereas the system of national accounts (SNA) "*measures what takes place in the* *economy, between which agents, and for what purpose*" (EC et. al. 2009:2), international accounts do the same for the processes occurring between an economy and the rest of the world, i.e. "*between residents of that economy and non-residents*" (IMF 2014:3).

Using a simplified linear communication model (Adler and Rodman 2006:12), we can say that based on rules given by reference manuals, statisticians "encode" collected data into aggregates that are recorded in accounts. This corresponds to "*activities and tasks to convert input data into statistical information*" (UNECE 2019:4). This statistical information will be "decoded" by the users of the statistical statement into knowledge on macroeconomic phenomena and processes. Depending on the target group's focus (see 4.1.2), the knowledge gained may be used "*for purposes of economic analysis, decision-taking and policymaking*" (EC et. al. 2009:1).

Of course, the communication that does take place is much more complex and includes feedback. The General Statistical Business Process Model begins with a phase called "*Specify Needs*" (UNECE 2019:11) where this feedback and the needs of users are considered. Nevertheless, a simplified model with "encoding" and "decoding" activities can be a useful approximation – especially because, as we will see, these activities correspond to different types of documents.

To sum up, we can say that the balance of payments is a *representation* of macroeconomic objects computed in statistical objects embedded in accounts.

4. Construction of a BOP corpus

4.1. Description

This research focuses on documents on the balance of payments published in three languages: English, French, and German. As mentioned, the BOP is embedded in an institutional framework: we have to do with official documents published by national organisations located in Austria, Belgium, Canada, France, Ireland, Germany, Luxembourg, Switzerland, the United Kingdom, the United States, and by international organisations (like the European Central Bank). We first present a typology of these organisations, then describe the different types of documents collected for the corpus under construction and last, present some metrics about the latter.

4.1.1. Typology of organisations producing documents on the BOP

The balance of payments is established either by a central bank (**CB**) or by a statistics office (**SO**). Reference documents are published mostly by other organisations (**OT**, e.g. the IMF) – although CBs and SOs may publish their own adapted guidelines that are equally considered as reference for their respective jurisdiction. Both CBs and SOs can be either a national (**N**) or an international (**I**) body. The organisations under scrutiny produce documents in English (**en**), in French (**fr**) and/or in German (**de**). Last, as mentioned in section 3, the covered jurisdiction can be either a country (**C**), an economic union (**Ec**) or a currency union (**Cr**). In some countries, the BOP is a common publication by the CB and the SO. This leads to four axes of analysis used for specific differentiation, as represented in TAB. 1.

Organisations	**Geographic competence**		**Language**[2]			**Type**			**Jurisdiction**		
	N	**I**	**de**	**en**	**fr**	**CB**	**SO**	**OT**	**C**[3]	**Ec**	**Cr**
Banque Centrale du Luxembourg[4]	•				•	•			LU		
Banque de France[4]	•				•	•			FR		
Bureau of Economic Analysis	•			•			•		US		
Bank of England	•			•		•			UK		
Bundesbank[4]	•		•			•			DE		
Central Statistics Office	•			•			•		IE		
CFMB		•		•				•	-	•	•
EC *et al.*		•		•				•	-	•	•
European Central Bank		•		•		•			-		•
Eurostat		•		•			•		-	•	
International Monetary Fund		•		•	•			•	-		

[2] ISO 639-1:2002 alpha-2 codes. Here, we only mention the official language(s) of the jurisdiction of the corresponding organisations (otherwise, the column "en" would be checked for all organisations).

[3] ISO 3166-1 alpha-2 codes.

[4] Central banks without official name in English.

Organisations	Geographic competence		Language[2]			Type			Jurisdiction		
	N	I	de	en	fr	CB	SO	OT	C[3]	Ec	Cr
National Bank of Belgium	•				•	•			BE		
Österreichische Nationalbank[4]	•		•			•			AT		
Office for National Statistics	•			•			•		UK		
Statistics Canada	•			•	•		•		CA		
Swiss National Bank	•		•		•	•			CH + FL		
Statec	•				•		•		LU		

TAB. 1 – *Typology of organisations producing documents on the BOP*

4.1.2. Typology of documents and possible target groups

The organisations mentioned in 4.1.1 produce and publish documents on the BOP on their respective websites. To construct our corpus, we have collected these documents and have asked experts who establish the BOP in these institutions to kindly cross-check whether any important document had been forgotten. Five categories have emerged from the analysis of the documents collected to construct the corpus. We have elaborated this typology based on pragmatic criteria, namely the communication situations (Maingueneau 2016:55) involved. This typology has been built using specific differentiation (see FIG. 2), similarly to the typology of organisations.

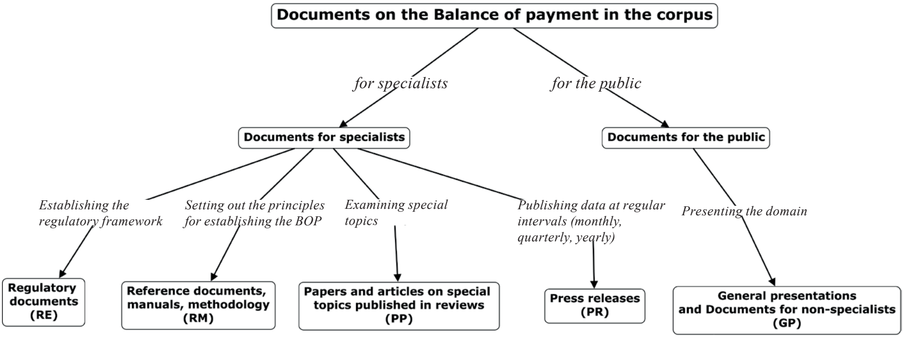

Documents on the Balance of payment in the corpus
for specialists
for the public
Documents for specialists
Documents for the public
Establishing the regulatory framework
Setting out the principles for establishing the BOP
Examining special topics
Publishing data at regular intervals (monthly, quarterly, yearly)
Presenting the domain
Regulatory documents (RE)
Reference documents, manuals, methodology (RM)
Papers and articles on special topics published in reviews (PP)
Press releases (PR)
General presentations and Documents for non-specialists (GP)

FIG. 2 – *Typology of documents in the BOP corpus*

In FIG. 2, we see that the corpus under construction can be subdivided using two axes of analysis: on the one hand, we distinguish the collected documents according to the knowledge necessary to understand them, which leads to two broad categories, documents for specialists vs documents for non-specialists; and on the other hand, we have analysed the content and the goals of these documents. We present hereafter this second axis of analysis, and the corresponding five categories of documents:

- We first have documents dealing with regulatory issues and laws for establishing the BOP, e.g. national laws stipulating that the BOP be established by the central bank. We group these documents under the label “Regulatory documents” (hereafter, **RE**).
- Secondly, we have documents presenting the principles of the domain e.g. statistical and methodological issues like data sampling, computing, conducting surveys, etc. These documents are gathered in the category “Reference documents, manuals, methodology” (hereafter, **RM**).
- Thirdly, we have collected documents dealing with specific research topics presented in reviews for specialists, which we have classified in the category “Research papers” (**PP**).
- The purpose of documents in our fourth category is to publish data on the balance of payments at regular intervals. They are grouped under the label “Press releases” (**PR**).

- Last, the fifth category is about "General presentations of the domain for non-specialists" (hereafter, **GP**).

In FIG. 3, we show an overview of the possible target groups of these documents, based on the focus of their respective activity: these are statisticians, scholars, investors and other actors in the financial markets, official bodies, like national governments, as well as the public. The underlying idea is that each of these target groups has specific needs when accessing to the knowledge expressed by documents on the BOP. This modelling has been validated by experts.

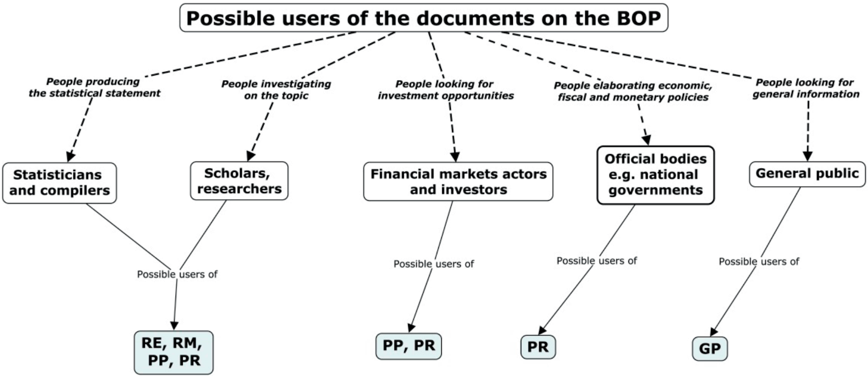

Possible users of the documents on the BOP
People producing the statistical statement
People investigating on the topic
People looking for investment opportunities
People elaborating economic, fiscal and monetary policies
People looking for general information
Statisticians and compilers
Scholars, researchers
Financial markets actors and investors
Official bodies e.g. national governments
General public
Possible users of
Possible users of
Possible users of
Possible users of
RE, RM, PP, PR
PP, PR
PR
GP

FIG. 3 – *Typology of possible target groups of BOP documents*

4.1.3. Metrics of the corpus under construction

The corpus currently contains 504 documents corresponding to almost 36 million characters. Its composition by language is strongly influenced by the presence of international reference documents in English (51%) and French (40%). Both are official languages of the United Nations, and also at the IMF – although the latter publishes most of its documents in English. However if we only consider documents produced by national organisations, the proportion of documents in German rises from 9% to 19%. But the percentage of documents in English also rises and equals 58% in the "national" corpus.

As for document types, the two main categories are reference documents, manuals and documents on methodology (RMs, 62 documents, almost 22 million characters) and press releases (PRs, 322 documents, nearly 10 million characters). As we will soon see, these two document types mostly correspond to specific activities performed with the balance of payments.

As far as the composition of the corpus by publication year is concerned, 99% of the documents have been published since 2009. This is fairly logical because the main RM currently in use (IMF 2009) was published in that year, and the whole corpus is determined by this conceptual framework. The very few exceptions correspond to regulatory documents (REs) on the BOP published before 2009 and still in force, as well as a few PPs.

Analysis of the corpus based on documents' *size* reveals a clear dominance of RMs since 2009. But if we consider the *number* of documents, we can distinguish two periods: documents produced before 2014 are mostly RMs and REs (published by international organisations), and since 2014, PRs (published by national organisations) prevail, with some RMs and PPs. The reason why most documents produced from 2009 to 2013 are RMs and REs is that the stress was put, in that period, on the definition of the conceptual and the regulatory framework based on IMF 2009. From 2014 onwards, the principles set out by IMF 2009 began to be applied by CBs and SOs producing PRs on the BOP[5]. This is the reason why PRs prevail since 2014. For details on these metrics, see Appendix 1.

4.2. Theoretical issues

The perspective taken by Tognini-Bonelli (2001) on corpus linguistics – as indeed by the majority of scholars in the field of corpus linguistics –, is the analysis of general language functions and features. Tognini-Bonelli (2001:2) defines a corpus as "*a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis*". However, it is as well possible to study specialised corpora in order to reach conclusions about the language used in specialised domains. As Flowerdew (2004:15) points out, specialised corpora are *"best suited in terms of their*

[5] Consequently, PRs published between 2009 and 2013 are based on the previous version of the RM (published in 1993) and are not compliant with IMF 2009. Logically, they will not be part of the corpus.

relevance for the purpose of understanding specific types of academic and professional language as general corpora may not be appropriate for this function on account of their internal composition". According to Tognini-Bonelli (2001:2), corpus linguistics "*operates within the framework of a contextual and functional theory of meaning"*. This feature is important if we want to base terminology work on a specialised corpus: it means we have to clearly distinguish between signified (object of study for linguistics) and concept (focus of terminology).

The BOP corpus is being built in order to help testify the terminology used in English, in French and in German, in the field of the balance of payments in a number of countries. It mostly consists of documents either explaining how to establish the BOP (RMs) or presenting the BOP as a statistical statement (PRs) for its users. RMs and PRs belong to a communication setting specialists-specialists (see 4.1.2) and can consequently be seen as *"the product of a community of controlled communication*" (Santos and Costa 2015:157); this setting does not provide much room for a free selection of lexical units based on contextual considerations: there is a prescriptive terminological framework, which reduces the range of possibilities. This is due to the fact that the domain is strongly standardised at the international as well as at the national and the regional level: the international framework set out by IMF 2009 and IMF 2014 applies to all IMF member states, and there are additional regional standards (e.g. OJEU 2012 and OJEU 2016, which apply to all member states of the European Union). Accordingly, the balance of payments established under these standards has to possess the same structure and the same headings. In other words, we do not expect to find a great variability at the lexical level[6].

We are now looking at the BOP corpus through the lens of the three different issues presented by Tognini-Bonelli (2001:55-62) as paramount when constructing or using a corpus: authenticity, sampling and representativeness.

4.2.1. Authenticity

This first criterion is actually not a real issue for our use of the BOP corpus for terminological work: we are obviously not going to create our cor-

[6] However, different lexical units may be used to denote the same concept in the different documents types (e.g. the term "components" in PRs and "functional categories" in RMs), based on the target groups. Similarly, variation will be observed in documents due to the presence of different varieties of English (or French or German).

pus “*in an artificial or experimental setting*” (Tognini-Bonelli 2001:55). As mentioned in section 4.1.2, we have collected the documents constituting the corpus directly from the websites of BOP authoring organisations.

4.2.2. Sampling

The main reference manual currently in use is the Balance of Payments Manual, 6th Edition, or BPM6 (IMF 2009). The IMF published the first edition in 1948. *“Each new edition of the Manual is introduced in response to economic and financial developments, changes in analytical interests, and accumulation of experience by compilers”* (IMF 2009, 1.17:3). IMF 2009 explores three main new fields: “*globalisation, increasing elaboration of balance sheet issues, and financial innovation*” (IMF 2009, 1.31:4). In order to cover these new fields, it introduces new concepts and new terms to denote them. This brings about either the use or the extension of use of terms that are rare or absent in the earlier editions, as well as the production of first definitions. This can be shown with the example of the term “special purpose entities”, abridged “SPEs”. This term denotes a concept corresponding to an object that is closely linked to a globalised economic environment and to financial innovation. Correspondingly, it is first mentioned (without definition) in the 5th edition of the Manual (IMF 1993), and a first definition – not yet based on an international consensus – is presented in IMF 2009 (4.50:58).

This example shows what Picht calls the *“life cycle”* of a concept (Picht 2013:11): this life cycle is “*the period of time a concept is considered correct and as such in active use*” (2013:14). In the case of the balance of payments, it is quite easy to determine the life cycle of concepts (and terms) that we have to focus on for sampling: central banks and statistics offices have been applying the principles set out by IMF 2009 since 2014. Concepts and categories have been changing since 2009 but no stable consensus has yet emerged among experts. The IMF will not publish the next edition of the Balance of Payments Manual before 2025[7]. This is why we take a synchronic perspective within the conceptual framework of IMF 2009. Consequently, our corpus encompasses – with the very few exceptions mentioned in 4.1.3 – documents published since 2009.

In order for a document to be selected, it must take place within the framework of IMF 2009 be it because it was produced after 2009 and explicitly

[7] It remains to be seen whether the future Manual (BPM7) will contain a definition of SPEs based on a large international consensus.

refers to IMF 2009 or because although produced before 2009 it is still valid or still in force (case of regulatory documents and some papers).

All documents included in the corpus are whole documents and not fragments: we agree with Bowker (1996:42-43) that this method is the best guarantee not to omit any portion containing valuable linguistic and terminological information.

The BOP corpus is open: this feature is especially relevant for documents of the category PRs, which are published at regular intervals by CBs and SOs, and will be collected directly from the websites and added to the corpus in the future.

4.2.3. Representativeness

We should notice with Rastier (2002) that no corpus can represent language as a whole, be it from a functional perspective – e.g. the object of linguistic description – or from a historical one – e.g. all documents available in a given language. Nevertheless, Rastier argues that it should be possible to assess the representativeness of a corpus based on a specific task or application for which we build the corpus[8]. But Tognini-Bonelli points out that "*at present we have no means of ensuring it [the representativeness], or even evaluating it objectively*" (2001:57).

As far as the BOP corpus is concerned, its authenticity and the methodology for sampling mentioned above (a synchronic perspective within the conceptual framework of IMF 2009) should make it representative of the discourse used in the domain of the balance of payments in the countries under scrutiny. Based on our sampling criteria, we have indeed collected documents that are typical for the domain. Its representativeness is thus qualitative, based on typicality and specialisation (Doualan 2018:6, § 17).

5. Role of the BOP corpus in ontoterminological modelling

We want to assess the role the BOP corpus can play in the elaboration of the ontoterminology of the balance of payments.

[8] «*Aucun corpus ne représente la langue: ni la langue fonctionnelle qui fait l'objet de la description linguistique, ni la langue historique, qui comprend l'ensemble des documents disponibles dans une langue. En revanche, un corpus est adéquat ou non à une tâche en fonction de laquelle on peut déterminer les critères de sa représentativité et de son homogénéité.*» (Rastier 2002 [online]).

5.1. Experts and corpora

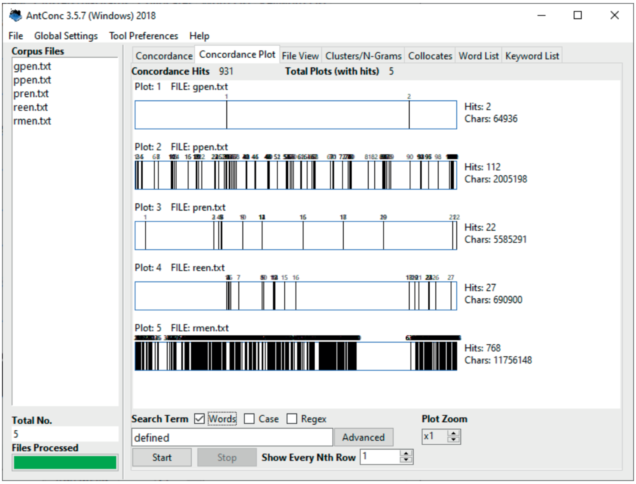
In terminology, we have to do with specialised domains, and the role of experts is therefore essential. This is, of course, true for this research. We want to build and model the concept system of the domain of the balance of payments in an ontology, and domain experts e.g. statisticians are best equipped with the necessary knowledge to help us do so. They can also give information about the use of terms.

As far as the BOP corpus is concerned, we can search it in order to extract term candidates and to testify the use of terms. Ideally, it will thus provide terms and equivalents in three languages for BOP concepts. Can it possibly provide more information? In other words, can it also help model the concept system? According to Santos and Costa, the linguistic perspective considers that texts are knowledge in action (2015:169), and "*the designation is, in discourse, a point of access to the concept*" (2015:156).

According to Condamines (2009:6), linguistic markers of relations will most probably be found in corpora consisting of texts intended to teach (e.g. reference manuals) because the authors of such documents often use contexts providing elements of definitions or at least knowledge rich contexts, aimed at helping non-expert users to understand a domain or a craft[9]. Based on our knowledge of BOP RMs, we agree with this position. A search of the whole corpus (RE, RM, PP, PR, GP) for the term "defined" with AntConc[10] shows (see FIG. 4) that almost 83% of the relevant contexts (768/931) are found in reference documents, manuals, and documents on methodology.

[9] "*De fait, il semble établi que les corpus les plus riches en marqueurs de relations et donc les plus susceptibles d'être utilisés pour construire des ontologie/terminologies sont les corpus de type didactique comme les « manuels ». En effet, les auteurs de manuels utilisent fréquemment des contextes définitoires ou du moins des contextes riches en connaissance destinés à faire comprendre un domaine ou une pratique pour des non-experts*" (Condamines, 2009:6).

[10] AntConc is a "*corpus analysis toolkit for concordancing and text analysis*" (cf. https://www.laurenceanthony.net/software/antconc/). All search results do not provide elements to build a definition, but some patterns (like "defined as") are very productive. We should notice that elements for definitions will also be provided by other patterns (e.g. "Ns like Ns", "understood as"…).



gpen.txt
ppen.txt
pren.txt
reen.txt
rmen.txt

Concordance Hits 931 Total Plots (with hits) 5

Plot: 1 FILE: gpen.txt
Hits: 2
Chars: 64936

Plot: 2 FILE: ppen.txt
Hits: 112
Chars: 2005198

Plot: 3 FILE: pren.txt
Hits: 22
Chars: 5585291

Plot: 4 FILE: reen.txt
Hits: 27
Chars: 690900

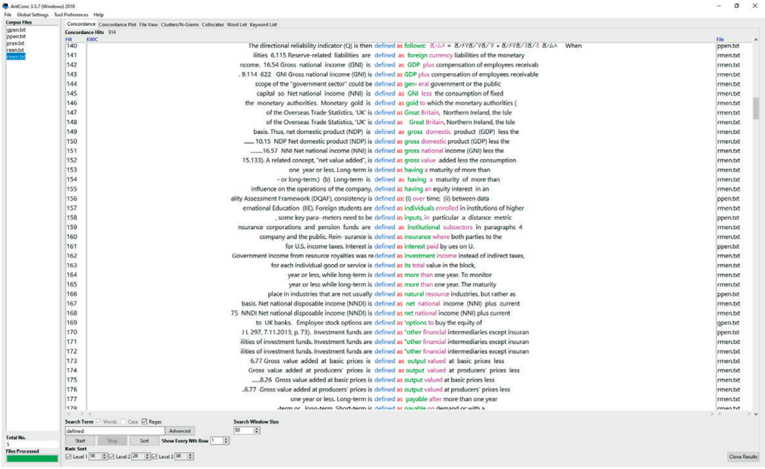
Plot: 5 FILE: rmen.txt
Hits: 768
Chars: 11756148

Total No.
5
Files Processed

Search Term ☑ Words ☐ Case ☐ Regex
defined
Show Every Nth Row 1
Plot Zoom x1

FIG. 4 – *Concordance plot for the search term "defined" in the five parts of the corpus (AntConc)*

The results displayed in the Concordance window (see FIG. 5) confirm the existence of numerous contexts with linguistic markers of relations in that type of documents. The same phenomenon can be observed in the French part of the corpus.


AntConc 3.5.7 (Windows) 2018
File Global Settings Tool Preferences Help
Corpus Files
Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List
Concordance Hits 914
Hit KWIC
File
Search Term ☑ Words □ Case ☑ Regex
Advanced
Search Window Size
Start Stop Sort Show Every Nth Row
Kwic Sort
Total No.
Files Processed
Clone Results

Corpus Files: gpen.txt, ppen.txt, pren.txt, reen.txt, rmen.txt

Concordance Hits 914

```
Hit  KWIC                                                                                                         File
140  The directional reliability indicator (Q) is then defined as follows: ð⁄ᐃ = ð⁄ᐟᐤð⁄ᐟᐤ⁄ᐟ + ð⁄ᐟᐤð⁄ᐟᐤ ð⁄ᐃ   When    ppen.txt
141  ilities 6.115 Reserve-related liabilities are defined as foreign currency liabilities of the monetary           rmen.txt
142  ncome. 16.54 Gross national income (GNI) is defined as GDP plus compensation of employees receivab              rmen.txt
143  . 9.114 622 GNI Gross national income (GNI) is defined as GDP plus compensation of employees receivable          rmen.txt
144  scope of the "government sector" could be defined as gen- eral government or the public                           rmen.txt
145  capital so Net national income (NNI) is defined as GNI less the consumption of fixed                              rmen.txt
146  the monetary authorities. Monetary gold is defined as gold to which the monetary authorities (                    rmen.txt
147  of the Overseas Trade Statistics, 'UK' is defined as Great Britain, Northern Ireland, the Isle                    rmen.txt
148  of the Overseas Trade Statistics, 'UK' is defined as Great Britain, Northern Ireland, the Isle                    rmen.txt
149  basis. Thus, net domestic product (NDP) is defined as gross domestic product (GDP) less the                       rmen.txt
150  ...... 10.15 NDP Net domestic product (NDP) is defined as gross domestic product (GDP) less the                   rmen.txt
151  ........16.57 NNI Net national income (NNI) is defined as gross national income (GNI) less the                   rmen.txt
152  15.133). A related concept, "net value added", is defined as gross value added less the consumption              rmen.txt
153  one year or less. Long-term is defined as having a maturity of more than                                          rmen.txt
154  - or long-term.) (b) Long-term is defined as having a maturity of more than                                       rmen.txt
155  influence on the operations of the company, defined as having an equity interest in an                            rmen.txt
156  ality Assessment Framework (DQAF), consistency is defined as: (i) over time; (ii) between data                    ppen.txt
157  ernational Education (IIE). Foreign students are defined as individuals enrolled in institutions of higher        rmen.txt
158  , some key para- meters need to be defined as inputs, in particular a distance metric                             ppen.txt
159  nsurance corporations and pension funds are defined as institutional subsectors in paragraphs 4                   rmen.txt
160  company and the public. Rein- surance is defined as insurance where both parties to the                           rmen.txt
161  for U.S. income taxes. Interest is defined as interest paid by ues on U.                                          ppen.txt
162  Government income from resource royalties was redefined as investment income instead of indirect taxes,           rmen.txt
163  for each individual good or service is defined as its total value in the block,                                   rmen.txt
164  year or less, while long-term is defined as more than one year. To monitor                                        rmen.txt
165  year or less while long-term is defined as more than one year. The maturity                                       rmen.txt
166  place in industries that are not usually defined as natural resource industries, but rather as                   ppen.txt
167  basis. Net national disposable income (NNDI) is defined as net national income (NNI) plus current                 rmen.txt
168  75 NNDI Net national disposable income (NNDI) is defined as net national income (NNI) plus current                rmen.txt
169  to UK banks. Employee stock options are defined as 'options to buy the equity of                                  gpen.txt
170  J L 297, 7.11.2013, p. 73). Investment funds are defined as "other financial intermediaries except insuran        ppen.txt
171  ilities of investment funds. Investment funds are defined as "other financial intermediaries except insuran       rmen.txt
172  ilities of investment funds. Investment funds are defined as "other financial intermediaries except insuran       rmen.txt
173  6.77 Gross value added at basic prices is defined as output valued at basic prices less                           rmen.txt
174  Gross value added at producers' prices is defined as output valued at producers' prices less                      rmen.txt
175  ......8.26 Gross value added at basic prices is defined as output valued at basic prices less                     rmen.txt
176  ..6.77 Gross value added at producers' prices is defined as output valued at producers' prices less               rmen.txt
177  one year or less. Long-term is defined as payable after more than one year                                        rmen.txt
178  -term or long-term. Short-term is defined as payable on demand or with a                                          rmen.txt
```

Search Term: defined
Search Window Size: 50
Show Every Nth Row: 1
Kwic Sort: ☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R
Total No.: 5

FIG. 5 – *Contexts with linguistic markers in the Concordance window (AntConc)*

That being said, when should we base our work on interactions with experts and when can we start from the corpus for building the concept system? In their research on the terminology of biological treatment of wastewaters, Santos and Costa first build a concept system with the help of experts, and at a second stage, they look for "*linguistic markers that could denote reformulation contexts*" in specialised texts (2015:171) and try to model the concept system based on texts. At that stage, they notice that the knowledge they have acquired by interacting with experts helps them to include in their modelling relations and concepts that do *not* have a linguistic expression in the corpus. Santos and Costa argue that there is "*a strong presence of the unsaid*" in specialised texts (2015:158). This aspect is also mentioned by Roche (2007b:52), who shows that building ontologies from texts is generally not satisfactory mainly because of the incompleteness of texts: *"understanding texts, and then understanding the meaning of terms, requires extra-linguistic knowledge which by definition is not included in the corpus*". In other words, if we solely rely on the corpus, we will not be able to identify all concepts of the domain, and we will possibly make errors in modelling. On the other hand, no expert sufficiently masters the entirety of his or her domain to be considered the only resource necessary for building the concept system. This is especially true for a domain like the balance of payments, which encompasses several specialities and activities[11].

5.2. A mixed approach

This leads us to adopt a mixed approach, "*using the onomasiological and semasiological points of departure of terminology whenever necessary*", as advocated by Santos and Costa (2015:177).

The author being responsible for the translation into French of documents related to the BOP at the Swiss central bank, he has acquired a semi-expertise on the domain – especially through interactions with experts. This is why the construction of the concept system will be corpus-based, e.g. based on RMs, and will also rely on that knowledge. In such a setting, semasiological and onomasiological approaches may be difficult to separate.

Building the system of concepts based on RMs relies on the following hypotheses: we should find a large proportion of terms pertaining to the "encoding" activities[12] in *reference documents, manuals*, and *regulatory documents*, and an important part of terms related to the "decoding" activities[13] in *press releases*. This assumption is quite consistent with the respective purpose and target groups of these documents types: by definition, reference manuals and regulatory documents give information to people establishing the BOP, they present the regulatory framework, explain how to conduct a survey, how to compile data into aggregates, etc. On the other hand, press releases mainly give information for people who do not establish the statistical statement but who are interested in the macroeconomic knowledge they can gain by "decoding" this statement: namely the users of the BOP[14].

To check the validity of our assumption on the respective contribution of the different document types for the linguistic expression of the activities performed within the balance of payments, we take the example of "institutional unit". Thanks to our semi-expertise, we already know that this is a term that denotes a concept corresponding to the basic entities observed by statisticians

[11] At the Swiss central bank, for example, not fewer than eighteen people carry different responsibilities related to the establishment of the balance of payments.

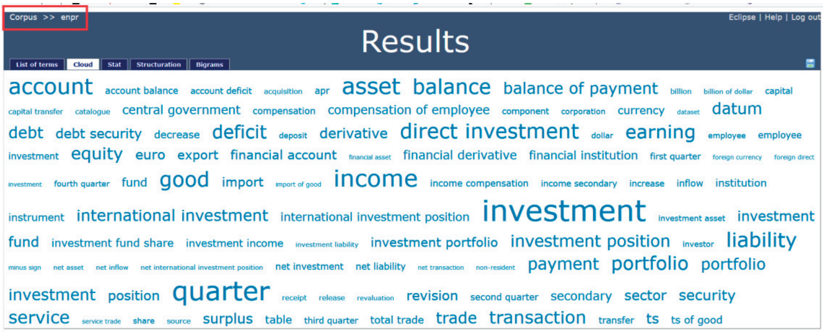
[12] Namely "*activities and tasks to convert input data into statistical information*".

[13] Namely activities to gain knowledge on macroeconomic phenomena "*for purposes of economic analysis, decision-taking and policymaking*".

[14] The contribution of the remaining text types (papers and general presentation) remains to be determined.

in the BOP. It is thus a central term in the “encoding” activity. It is defined as an economic entity possessing four essential characteristics, whose definition in natural language formed after IMF 2009 reads “*Economic entity entitled to own goods or assets in its own right, able to take economic decisions, able to enter into contracts and with a complete set of accounts or the possibility thereof*” (Carsenty 2020:89-91).

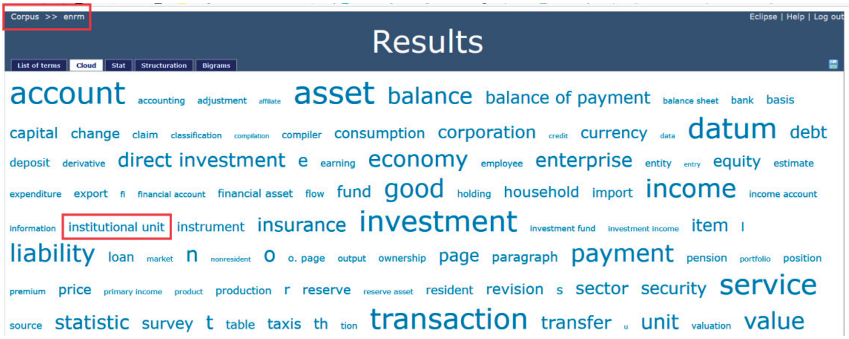
Let us now put ourselves in the shoes of a user who does not know that “institutional unit” is a term. Analysing the different document types for term candidates will help him or her include it in the list of term candidates … or not. If s/he analyses exclusively press releases (see FIG. 6), this term does not appear at all as a term candidate in the cloud displayed in TermoStat[15]. It only appears in that cloud if s/he explores reference documents and manuals (see FIG. 7). Opening the Contexts (concordances) tab in TermoStat (see FIG. 8) will provide the user with knowledge rich contexts contained in reference documents and manuals, allowing the construction of a definition in natural language which should be similar to the one we have mentioned in the previous paragraph. The same phenomenon can be observed when analysing the French part of the corpus.



account account balance account deficit acquisition apr asset balance balance of payment billion billion of dollar capital capital transfer catalogue central government compensation compensation of employee component corporation currency dataset datum debt debt security decrease deficit deposit derivative direct investment dollar earning employee employee investment equity euro export financial account financial asset financial derivative financial institution first quarter foreign currency foreign direct investment fourth quarter fund good import import of good income income compensation income secondary increase inflow institution instrument international investment international investment position investment investment asset investment fund investment fund share investment income investment liability investment portfolio investment position investor liability minus sign net asset net inflow net international investment position net investment net liability net transaction non-resident payment portfolio portfolio investment position quarter receipt release revaluation revision second quarter secondary sector security service service trade share source surplus table third quarter total trade trade transaction transfer ts ts of good

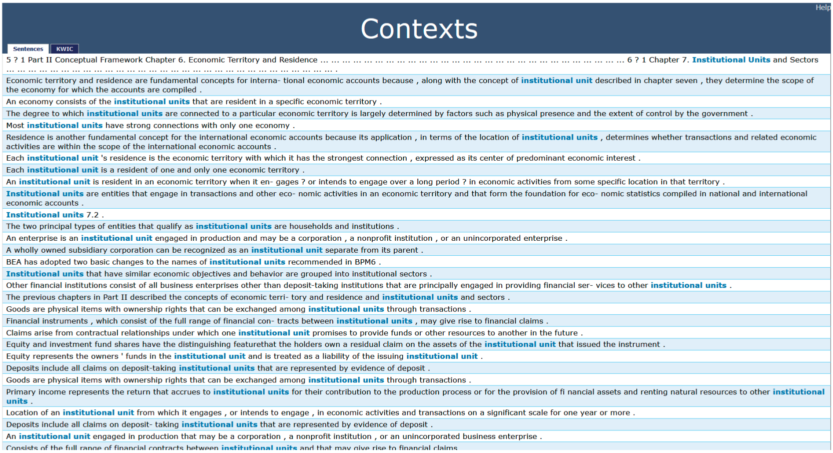
FIG. 6 – *Cloud of term candidates extracted from press releases in English (corpus >>enpr) displayed in TermoStat*

[15] A tool for automatic term acquisition developed by Patrick Drouin at the University of Montréal, Canada (http://termostat.ling.umontreal.ca/).



account accounting adjustment affiliate asset balance balance of payment balance sheet bank basis
capital change claim classification compilation compiler consumption corporation credit currency data datum debt
deposit derivative direct investment e earning economy employee enterprise entity entry equity estimate
expenditure export fi financial account financial asset flow fund good holding household import income income account
information institutional unit instrument insurance investment investment fund investment income item I
liability loan market n nonresident O o. page output ownership page paragraph payment pension portfolio position
premium price primary income product production r reserve reserve asset resident revision s sector security service
source statistic survey t table taxis th tion transaction transfer u unit valuation value

FIG. 7 – *Cloud of term candidates extracted from reference documents and manuals in English (corpus >> enrm) displayed in TermoStat*



5 ? 1 Part II Conceptual Framework Chapter 6. Economic Territory and Residence … 6 ? 1 Chapter 7. **Institutional Units** and Sectors … .

Economic territory and residence are fundamental concepts for interna- tional economic accounts because , along with the concept of **institutional unit** described in chapter seven , they determine the scope of the economy for which the accounts are compiled .

An economy consists of the **institutional units** that are resident in a specific economic territory .

The degree to which **institutional units** are connected to a particular economic territory is largely determined by factors such as physical presence and the extent of control by the government .

Most **institutional units** have strong connections with only one economy .

Residence is another fundamental concept for the international economic accounts because its application , in terms of the location of **institutional units** , determines whether transactions and related economic activities are within the scope of the international economic accounts .

Each **institutional unit** 's residence is the economic territory with which it has the strongest connection , expressed as its center of predominant economic interest .

Each **institutional unit** is a resident of one and only one economic territory .

An **institutional unit** is resident in an economic territory when it en- gages ? or intends to engage over a long period ? in economic activities from some specific location in that territory .

Institutional units are entities that engage in transactions and other eco- nomic activities in an economic territory and that form the foundation for eco- nomic statistics compiled in national and international economic accounts .

Institutional units 7.2 .

The two principal types of entities that qualify as **institutional units** are households and institutions .

An enterprise is an **institutional unit** engaged in production and may be a corporation , a nonprofit institution , or an unincorporated enterprise .

A wholly owned subsidiary corporation can be recognized as an **institutional unit** separate from its parent .

BEA has adopted two basic changes to the names of **institutional units** recommended in BPM6 .

Institutional units that have similar economic objectives and behavior are grouped into institutional sectors .

Other financial institutions consist of all business enterprises other than deposit-taking institutions that are principally engaged in providing financial ser- vices to other **institutional units** .

The previous chapters in Part II described the concepts of economic terri- tory and residence and **institutional units** and sectors .

Goods are physical items with ownership rights that can be exchanged among **institutional units** through transactions .

Financial instruments , which consist of the full range of financial con- tracts between **institutional units** , may give rise to financial claims .

Claims arise from contractual relationships under which one **institutional unit** promises to provide funds or other resources to another in the future .

Equity and investment fund shares have the distinguishing featurethat the holders own a residual claim on the assets of the **institutional unit** that issued the instrument .

Equity represents the owners ' funds in the **institutional unit** and is treated as a liability of the issuing **institutional unit** .

Deposits include all claims on deposit-taking **institutional units** that are represented by evidence of deposit .

Goods are physical items with ownership rights that can be exchanged among **institutional units** through transactions .

Primary income represents the return that accrues to **institutional units** for their contribution to the production process or for the provision of fi nancial assets and renting natural resources to other **institutional units** .

Location of an **institutional unit** from which it engages , or intends to engage , in economic activities and transactions on a significant scale for one year or more .

Deposits include all claims on deposit- taking **institutional units** that are represented by evidence of deposit .

An **institutional unit** engaged in production that may be a corporation , a nonprofit institution , or an unincorporated business enterprise .

Consists of the full range of financial contracts between **institutional units** and that may give rise to financial claims

FIG. 8 – *Contexts (concordances) with the term candidate "institutional units", extracted from reference documents and manuals, displayed in TermoStat*

6. Conclusion and future work

Constructing a text corpus for a work on the ontoterminology of the balance of payments leads to develop a typology of texts included in that corpus. This typology is based on the following assumption: not all texts will deliver the same kind of information for the ontoterminology. As a matter of fact, in terminology, we both need to know the portion of the world which LSP texts deal with and to access the system of concepts experts have built to think about that portion of the world. This is why we have elaborated our typology of texts in the BOP corpus based on the communication situations (Maingueneau 2016:55) involved by these texts. These situations are characterised by the identity of the agents, the purpose and the general setting (Charaudeau 2009: 50).

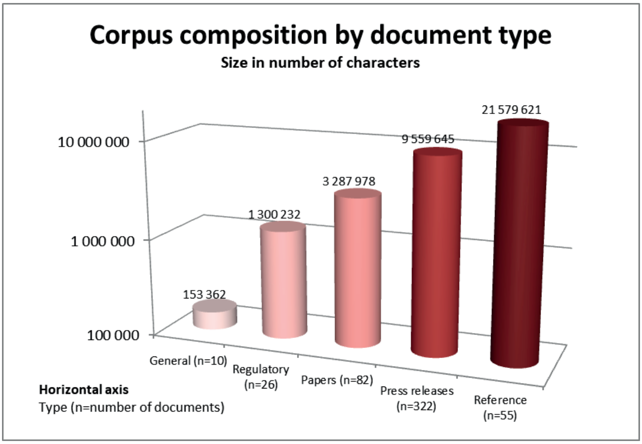
The construction of a corpus in the domain of the balance of payments gives different insights. Working on this corpus will help testify terms in use in that domain. An important proportion of the BOP corpus consists of reference documents, manuals and texts on methodology. These documents contain numerous knowledge rich contexts with linguistic markers of relationships which will help build the concept system in our terminology. On the other hand, experts remain essential, at both the modelling and the validation stage. We have constructed the BOP corpus for the purpose of terminological analysis and we possess domain knowledge because we translate documents that have been collected for the corpus. The characteristics of that corpus, our preliminary knowledge on the balance of payments and the possibility to have access to experts thus lead to a mixed approach including both semasiological and onomasiological stages.

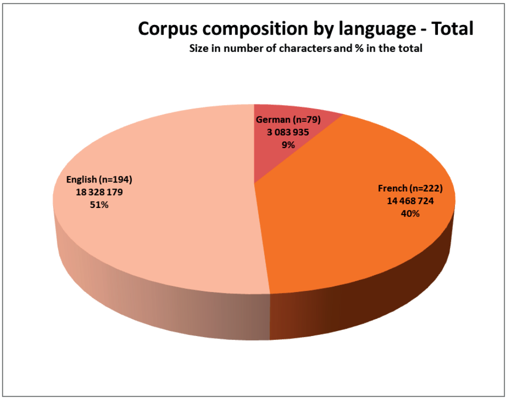
Now that the BOP corpus exists, the next stage will be to model the system of concepts of the balance of payments, and corresponding terms in English, in French and in German. Thanks to statisticians who establish Switzerland's balance of payments, we have contact with their counterparts in other institutions. We are glad to have them for this research.

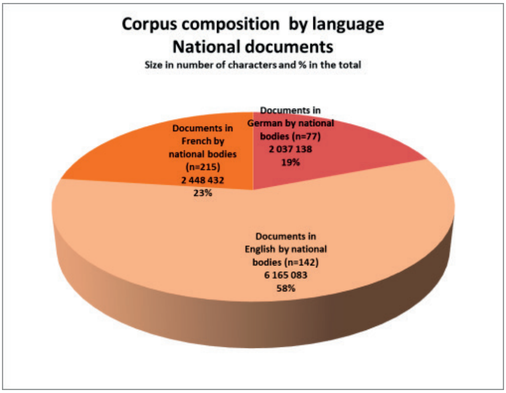
Abbreviations used

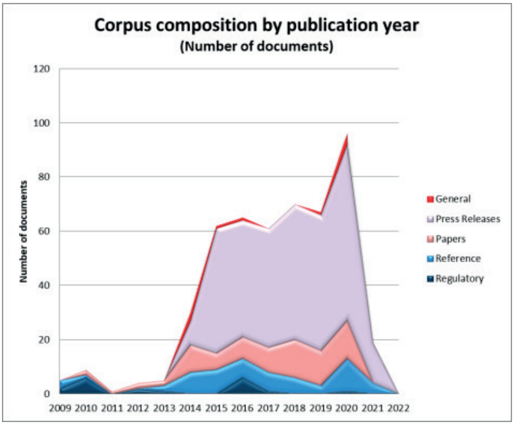
BOP	Balance of payments
CBs	Central banks
CFMB	Committee on Monetary, Financial and Balance of Payments Statistics
EC	European Community
GPs	General presentations
IMF	International Monetary Fund
PPs	Papers
PRs	Press releases
REs	Regulatory documents
RMs	Reference documents, manuals, methodology
SOs	Statistics offices

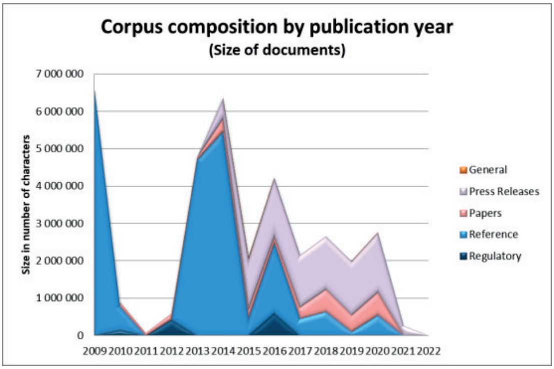
Appendix 1: Metrics on the corpus




Corpus composition by language - Total
Size in number of characters and % in the total
German (n=79) 3 083 935 9%
French (n=222) 14 468 724 40%
English (n=194) 18 328 179 51%


Corpus composition by language National documents
Size in number of characters and % in the total
Documents in German by national bodies (n=77) 2 037 138 19%
Documents in French by national bodies (n=215) 2 448 432 23%
Documents in English by national bodies (n=142) 6 165 083 58%


Corpus composition by publication year
(Number of documents)
Number of documents
0
20
40
60
80
100
120
2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022
General
Press Releases
Papers
Reference
Regulatory



References

Adler, R. B. & G. Rodman. 2006. “Understanding Human Communication, Ninth Edition”, Oxford University Press, Oxford, New York.

Bowker, L. 1996. “Towards a corpus-based approach to terminography” in Terminology 3:1 (1996), pp. 27-52, John Benjamins, Amsterdam. DOI 10.1075/term.3.1.03bow.

Carsenty, S. 2020. « Première approche de la modélisation ontoterminologique de la balance des paiements » in Terminologie & Ontologie: Théories et Applications. Actes de la conférence TOTh 2020, Université Savoie Mont Blanc, Chambéry, 26 & 27 novembre 2020, pp. 81-99.

Charaudeau, P. 2009. « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique », Corpus, 8, « Corpus de textes, textes en corpus », Nice, France, pp. 37-66.

Condamines, A. 2009. « Comment peut-on construire une ontologie personnelle à partir de textes? Considérations linguistiques », TIA’09, Actes de la 8^{e} conférence internationale Terminologie et Intelligence Artificielle, 18-20 novembre 2009, Toulouse, France (online: https://www.irit.fr/TIA09/thekey/articles/condamines.pdf).

Doualan, G. 2018. « De la représentativité à la spécialisation: exemple d’un petit corpus sur la synonymie », Corpus, 18 | 2018 (online: http://journals.openedition.org/corpus/3331).

Flowerdew L. 2004. “The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings’, in U. Connor and T. Upton (eds) Discourse in the Professions: Perspectives from Corpus Linguistics. Amsterdam: John Benjamins, pp. 11-33.

International Organization for Standardization (ISO). 2009. ISO 704:2009 “Terminology work — Principles and methods”. Geneva.

International Organization for Standardization (ISO). 2019. ISO 1087-2019 “Terminology work and terminology science — Vocabulary”. Geneva.

Maingueneau, D. 2016. « Analyser les textes de communication », Armand Colin, Paris.

Picht, H. 2013. “Concepts as Reflection of Societal Changes” in “Terminologija” | 2013 | 20, pp. 10-23 (Vilnius, Lithuania).

Rastier, F. 2002. « Enjeux épistémologiques de la linguistique de corpus ». Texto! [en ligne], juin 2004. Rubrique Dits et inédits. (online: http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html).

Roche, C. 2007a. « Le terme et le concept: fondements d’une ontoterminologie », Actes de la conférence TOTh 2007: « Terminologie & Ontologie: Théories et Applications » - Annecy, France, 1er juin 2007, pp. 1-22.

Roche, C. 2007b. “Saying is not modelling”. 4th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2007), June 2007, Funchal, Portugal. pp. 47-56.

Roche, C. 2012. Ontoterminology: How to unify terminology and ontology into a single paradigm. In: Proceedings of LREC 2012, Eighth International Conference on Language Resources and Evaluation, May 2012, Istanbul, Turkey, pp. 2626-2630.

Roche, C. & M. Papadopoulou. 2020. « Rencontre entre une philologue et un terminologue au pays des ontologies », Revue Ouverte d’Intelligence Artificielle, Volume 1, n° 1 (2020), pp. 43-70.

Santos, C. & R. Costa. 2015. “Domain specificity. Semasiological and onomasiological knowledge representation” in “Handbook of Terminology”, Volume 1, edited by Hendrik J. Kockaert and Frieda Steurs, John Benjamins, Amsterdam, pp. 153-179.

Sowa, J. 2000. “Knowledge Representation - Logical, Philosophical, and Computational Foundations”, Brooks/Cole, Pacific Grove, USA.

Tognini-Bonelli, E. 2001. “Corpus Linguistics at Work”. John Benjamins Publishing Company Amsterdam / Philadelphia (USA).

BOP and International accounts (excerpts from the corpus)

European Communities, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations and World Bank (EC *et al.*). 2009. “System of National Accounts 2008 (SNA 2008)”, United Nations, New York.

International Monetary Fund (IMF). 1993. “Balance of Payments Manual”, 5th Edition, Washington, D.C.

International Monetary Fund (IMF). 2009. “Balance of Payments and International Investment Position Manual”, 6th Edition (BPM6), Washington, D.C.

International Monetary Fund (IMF). 2014. Balance of payments and international investment position compilation guide. – Washington, D.C.

United Nations Economic Commission for Europe (UNECE) 2019. Generic Statistical Business Process Model V.5.1 (GSBPM)

Official Journal of the European Union (OJEU). 2012. L 166/22. COMMISSION REGULATION (EU) No 555/2012 of 22 June 2012 amending Regulation (EC) No 184/2005 of the European Parliament and of the Council on Community statistics concerning balance of payments, international trade in services and foreign direct investment, as regards the update of data requirements and definitions

Official Journal of the European Union (OJEU). 2016. L171/144. REGULATION (EU) 2016/1013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 8 June 2016 amending Regulation (EC) No 184/2005 on Community statistics concerning balance of payments, international trade in services and foreign direct investment

Résumé

La modélisation ontoterminologique d’un domaine relève d’une approche résolument conceptuelle de la terminologie. Cette approche accorde un rôle central aux experts tant pour l’élaboration du système de concepts que pour la validation des termes du domaine. Il n’en demeure pas moins que l’ontoterminologie peut tirer profit de l’utilisation de corpus. Cet article présente les questions posées par la construction d’un corpus de textes dans un domaine socialement construit et dont les objets sont immatériels, à savoir la balance des paiements, et s’efforce de déterminer le rôle que ce corpus peut jouer dans l’élaboration de l’ontoterminologie de ce domaine.

Prix jeunes chercheurs

Collaborative terminology management in a business environment: a case study in the field of wood paints and coatings

Cristina Farroni*

*University of Macerata
Dipartimento di Studi Umanistici
c.farroni2@unimc.it

Abstract. Companies operating at an international level often must keep pace with large translation volumes as well as with related multilingual terminology. Their terminology is deeply influenced by daily practice, as well as by the communities of users who deal with the terminology intralingually as well interlingually. This contribution aims to analyse terminological issues in a specific case study, i.e., an Italian company that sells wood paints and coatings. Starting from some initial findings of a quantitative analysis carried out in the Italian region of the Marches we go on to describe terminological problems as they occur in a real translation workflow, as observed within a local company. The aim is to present these issues by considering each department involved in the workflow and to outline the collaborative and interdisciplinary analysis implemented to disambiguate variants and enhance the quality of translations.

1. Introduction

According to ISO 1087 (2019), terminology is a “set of designations and concepts belonging to one domain or subject.” Every domain is characterised by a certain language for special purposes, namely a “natural language used in communication between experts in a domain and characterized by the use of specific linguistic means of expression” (ISO 1087: 2019). A natural language is, in turn, a “language that is or was in active use in a community of people, and the rules of which are mainly deduced from usage” (ISO 1087: 2019). When it comes to organizations and businesses, communities and usage play a key role. Even when comparing terminology used by companies active in the same field, there are asymmetries and deviations from shared standards[1]. In some cases, companies intentionally differentiate names of products and services from those used by competitors[2], whereas, in other cases, terminological deviations occur unintentionally and are heavily influenced by cognitive processes, as well as by internal practices known to employees but not to external stakeholders. Bertaccini and Lecci (2010) explain that companies can be seen as microcosms partially isolated from the outside world where oral communication mainly occurs among people within the company. The output of oral communication leads to shared practices that enter written texts over time and give rise to local terminological preferences and variations. This in turn has an impact on the transfer of terms interlingually.

The advantages of terminology management are presented in several studies that empirically show how to collect, store, and manage multilingual terminology (Drewer, Schmitz 2017). Schmitz and Straub (2016, 24) also outline typical consequences of terminological problems in businesses that lead to lower-quality translations (e.g., mistakes in technical documentation and/or translations) as well as to higher costs and waste of time. In the relevant literature, many case studies of terminology management are referred to companies located in Switzerland (for instance, Krämer 2020, Holdermann 2020) – where plurilingualism plays a key role – and in Germany (for instance, Barsch-Harjau 2020), where Drewer (2019, 66) observes an increase in job advertisements addressed to terminologists. On the contrary, no significant case studies of companies dealing internally with multilingual terminology

[1] This is also the reason why de Vecchi (2016) interestingly suggests reconsidering the notion of domain. He argues that a knowledge domain should further include "*domaines de activité*" and, more specifically, "*domaines d'exploitation*". Such a distinction would also mirror different company-speaks and terminological practices shared by employees of a given company but not by other companies in the same field. Furthermore, a specific company can be seen as a group of communities, where some terms are shared by the whole company, while other terms may be known only to a certain community within the company working on a given project (cfr. de Vecchi 2012, 79). This situation sometimes leads to terminological asymmetries within the company itself.

[2] de Vecchi (2020, 243) uses the example of burger names chosen by McDonald's and Burger King and suggests that these should not be treated just as proper names because they are actively used by employees, enter their professional discourse and mirror the specific and unique reality of each company. This is why, as far as companies are concerned, de Vecchi suggests a pragmaterminological approach (Delavigne and de Vecchi 2016; de Vecchi 2020) where terminology is seen as part of company-speak and is deeply influenced by the community of people working at the company and by the knowledge they share.

have been identified in literature related to the Italian landscape. Within the framework of a Ph.D. project carried out at the University of Macerata, we decided to carry out a more in-depth analysis to find out whether local companies had experienced issues related to multilingual terminology. Results of a first quantitative analysis will be briefly outlined in this contribution and a specific case-study will be presented. The aim is to show different types of terminological issues occurring at the observed company and in different departments. To solve these issues and enhance the quality of translations at the observed company, we then carried out a collaborative analysis of terminology involving technicians as well as experts in other disciplines (such as SEO and marketing). The aim was also to provide the company's employees with a work method that could allow them in future to deal with terminological issues.

2. Quantitative analysis

From October 2020 to January 2021, we asked SMEs and large enterprises to respond to a questionnaire that covered four thematic areas: general information (size, sector, foreign markets), language management within the companies (linguistic competences, departments involved in communication with foreign partners, language courses for employees), translation workflows (internal or external, employees and stakeholders involved) and translation-related issues[3]. Within the scope of this contribution, we will focus only on the fourth thematic area before moving on to the case study.

Questionnaires were sent both to SMEs and to larger enterprises and, more specifically, to employees typically involved in interlingual communication, such as those in export departments. As far as SMEs are concerned, it was not always possible to identify a specific department dedicated to export and translation tasks. Therefore, questionnaires were sent to relevant employees of other departments, such as marketing or top management. We were able to collect answers from 205 local companies and gained a first overview of their translation and terminology-related issues. More specifically, we asked respondents whether they had ever experienced one or more of the following problems:

1) mistakes in translations carried out by external translators or agencies[4],

[3] A more in-depth presentation of these results will be part of the final Ph.D. thesis.

[4] As some researchers in the field of corporate language management point out (Welch *et al*. 2001, Lesk *et al*. 2017, 279), translation issues are sometimes caused by external trans-

2) mistakes in translations carried out internally,
3) terminological variants in translations over time and/or,
4) different terms used by different departments within the company to designate the same products[5].

As shown below, only 26% of the companies did not experience any issue; 3% of respondents did not know and 71% of the companies experienced one or more of the above-mentioned issues.

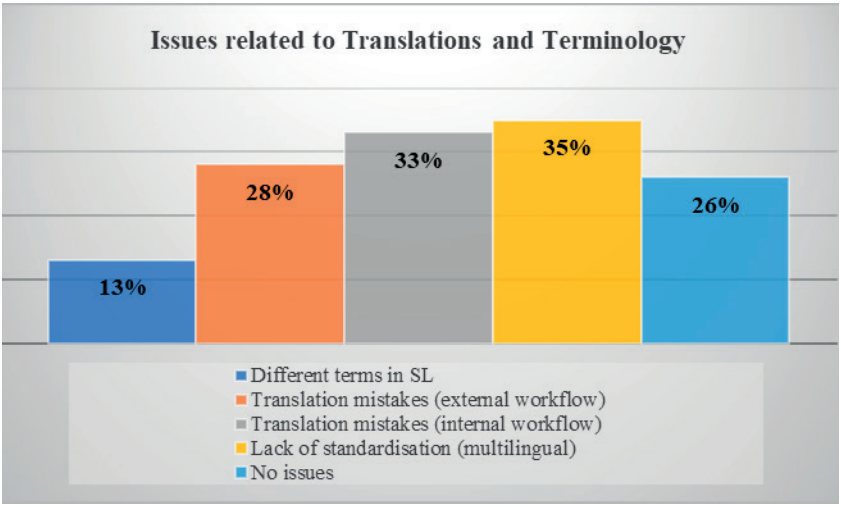


FIG. 1– *Issues related to translations and terminology in local companies.*

Some major terminological issues are related to the lack of standardisation of multilingual documentation and to translation mistakes; 35% of respondents claimed a lack of standardisation in their multilingual terminology, while 33% of respondents identified mistakes in translations carried out internally and 28% of respondents stated that these mistakes had been made by external translators. Finally, 13% of respondents experienced a lack of standardisation intralingually as well.

lators' or translation agencies' lack of specialized knowledge and terminology, possibly resulting in inaccurate translations.

[5] Some common terminological issues are indicated in Schmitz and Straub (2016, 20-24).

Some of these terminological issues have been further analysed in a local company. At the company, data have been collected through participant observation from October to December 2020. By doing so, it was possible to observe the translation workflow and identify some causes of terminological issues. At a later stage we cooperated with employees to improve the overall quality of the company's terminology.

3. Case study: department-related terminological issues and collaborative improvement

The company involved in the project produces paints and coatings for wood and has seven branches abroad. Documentation related to products and services must be translated into at least six target languages – English, German, Spanish, Polish, French, and Chinese – to reach customers and business partners abroad. The company's translation workflow is both internal and external. In fact, technical data sheets are drawn up by the product development department and then translated internally by the export department. Catalogues and newsletters are edited by the communications and marketing department and then are translated externally by a translation agency. At least three departments are involved in the translation workflow: the product development department, the export department and the communications and marketing department.

One of the most interesting findings was, therefore, the identification of department-related terminological issues that involved different text types intralingually as well as interlingually. These issues as well as their causes will be outlined in the following sections, focusing mainly on product designations in Italian, English, and German.

3.1. Intralingual terminological issues: product development department

A first analysis was based on terminology extracted from a corpus of company texts. We collected Technical Data Sheets and Technical Manuals written by technicians and extracted 250 keywords, automatically as well as manually. We then compared their use in the company's documentation with definitions provided in manuals and general standards (mainly ISO and DIN) of the paint industry. This comparison, as well as dialogue with technicians, allowed us to identify some deviations from shared and well-known standards and to identify the causes of some of the translation mistakes experienced by the company. In the following section, these terminological asymmetries will be described by providing some meaningful examples and by focusing on terminological issues in the source language – Italian – as well as on the related consequences in two target languages: English and German.

One of the main translation mistakes was related to the improper use of the term *colorante* in the source language. Within the company, this term is used to refer to two different kinds of products in the paint industry:

1) the first usually is referred to as *dye* or *dyestuff* in English. According to ISO 4618:2014, a *dyestuff* is a specific type of colouring material that "is soluble in the application medium";

2) the second is expressed by the English term *wood stain*. According to ISO 4618:2014, wood stain is a "penetrating composition containing a dyestuff that changes the colour of a wood surface, usually transparent and leaving no surface film, the solvent for which may be oil, denaturated alcohol or water."

In practice, technicians usually refer to these two different concepts as *colorante* in Italian. A more accurate distinction would require them to use the term *tinta* (instead of *colorante*) when talking about wood stains and *colorante* when referring to the soluble substances contained in the wood stain. From a terminological point of view, *colorante* and *tinta* refer to two different types of products sold by the company and cannot be considered synonyms. The concepts they designate do not have the same intension and, rather are connected by a part-whole relation (*colorante*, i.e., dyestuff, is part of *tinta*, i.e., wood stain). A possible reason why technicians tend to use the term *colorante* when referring to wood stains, is that this term – if compared to the correct term *tinta* – also reflects quite well the function of a wood stain, (i.e., 'colouring' the wood) and has therefore been widely used over time. As far as translations are concerned, this terminological inaccuracy led to some wrong translations in English as well as in German. Translators, working for a translation agency as well as those within the company, are not always aware of this distinction and sometimes lack enough context to provide the correct translation.

Another main terminological issue arises from terminological practices shared by the community of experts within the company but that, at the same time, represent a deviation from the accepted and known terminological standards of the community of experts. This stands out when analysing the Italian terms *induritore* and *catalizzatore*. The term *catalizzatore* (i.e., catalyst) is widely used in chemistry to identify substances that accelerate chemical reactions but do not take part in the reaction itself. Nevertheless, when referring to so-called *two-component coatings* in the paint industry, the right term should be *induritore* (i.e., *hardener*): in fact, in this specific case, the substance takes part in the reaction and becomes a solid component of the coating product. Even though experts are aware of the distinction, they admitted that the term *catalizzatore* often is used – incorrectly – instead of *induritore* to address a heterogeneous target audience. However, this practice only considers the use of the term *catalizzatore* in the domain of chemistry in general and does not account for the peculiarities of terminology related to two-component coatings and lead to incorrect terminology in target languages as well.

One last issue is represented by multi-word terms where one element remains implicit. The following table shows the extended designations for some painting products (first column), the terms used at the company (second column) and the cognitive focus involved (third column).

Multi-word terms	**Term frequently used at the company**	**Focus**
Impregnante di finitura Head: *impregnante* (product category) Modifier: *di finitura* (product function) Impregnante biocida Head: *impregnante* (product category) Modifier: *biocida* (product function)	Impregnante	Focus on the main category of products (*impregnante*) but not on its specific function (*biocida/ di finitura*).

Multi-word terms	Term frequently used at the company	Focus
Additivo uniformante Head: *additive* (product category) Modifier: *uniformante* (product function) Fondo uniformante Head: *fondo* (product category) Modifier: *uniformante* (product function)	Uniformante	Focus on the function (*uniformante*) but not on the product category (*additivo*/*fondo*).

Tab. 1 – *Designations of painting products and terms used at the company where one element remains implicit.*

As far as the first example is concerned, the head *impregnante* (i.e., a stain for outdoor wood) refers to one of the main categories of products sold by the company, while the modifiers *di finitura* and *biocida* refer to different functions performed by this category of products and correspond to two different product types (the first one – *impregnate di finitura* – is used as a final layer with a finishing function, while the second one – *impregnante biocida* – provides a protection against harmful organisms in wood). Technicians at the company often use the generic term *impregnante*, without providing any further information about its function because they know that the company produces *impregnanti di finitura* and therefore do not need a more accurate distinction. This implicit knowledge is shared by technicians but not with external translators, thus making the choice of the correct target term more difficult. On the contrary, the second example shows how technicians refer – in the oral as well as written language – to the function of certain products without providing any information about the category of the product involved. The modifier *uniformante* (i.e., *uniforming* or *levelling*) may be used in connection with the heads *additivo* (*additive*) or *fondo* (*base coat*, i.e., the first layer of a coating cycle) and refers to specific functions performed by these two categories of products. The multi-word term *additivo uniformante* refers to additives that, when added to a coating material, make it more homogeneous and easier to apply. On the contrary, the multi-word term *fondo uniformante* refers to coating materials used to fill the gaps on wooden surfaces and enhance the application of other layers of coating. We identified occurrences of *uniformante* used as a single-word term in instructions and on lists of product names. However, in target languages such as German and English, it is not possible to translate the term *uniformante* alone and it is necessary to understand not only the function but also the category of the product involved.

The three examples described above reveal different causes of terminological issues at the observed company. In some cases, oral practices shared by employees tend to settle and be transferred to the written language. In other cases, terminological issues are caused by a lack of details or by implicit knowledge not shared with people involved in the translation workflow.

In the following section, we will consider issues related to the identification of equivalent terms in target languages (English and German) and the need to compare the concepts they refer to in different countries along with manufacturing practices in these countries.

3.2. Interlingual terminological issues: export and marketing departments

As far as the translation of technical data sheets is concerned, the second company department involved is the export department. In this case, employees must face a further terminological issue: identifying the correct target terms for concepts related to manufacturing practices and standards that differ from one country to another. For instance, in the paint industry, the term *isolante* refers to a base coat for wood that prevents substances and components of the substrate from migrating to the surface. The proper term in English is *sealer*, but one must be careful since the term *sealer* also refers, in some countries such as the UK, to a specific type of transparent finish.

Another example is the above-mentioned term *impregnante*. In Italian, *impregnante* refers to a specific type of protective coating for wooden furniture placed outdoors that may contain biocides and fungicides and sometimes is used as a finishing topcoat. The term *impregnante* has been translated into German for a long time by employees with the terms *Imprägnierung* or *Imprägniermittel*. Even though these two terms seem, at first glance, to be the most suitable translation, further analysis was needed. Comparison of the main features of the product category *impregnante* – as meant and sold by the company – with *Imprägnierungen* – as used in the German market – reveals the following differences:

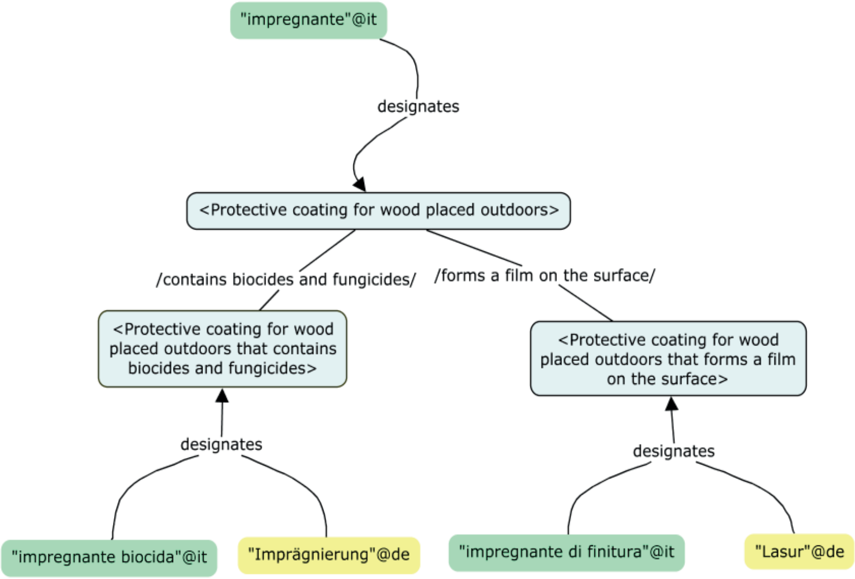

"impregnante"@it
designates
<Protective coating for wood placed outdoors>
/contains biocides and fungicides/
/forms a film on the surface/
<Protective coating for wood placed outdoors that contains biocides and fungicides>
<Protective coating for wood placed outdoors that forms a film on the surface>
designates
designates
"impregnante biocida"@it
"Imprägnierung"@de
"impregnante di finitura"@it
"Lasur"@de

FIG. 2 – *Comparison of the term impregnante (IT) and its equivalents in German.*

One can observe that there is only a partial equivalence between what Italian experts call *impregnante* and what German technicians mean by *Imprägnierung.* A so-called *Impregnante*, in Italy, does not necessarily contain biocides and fungicides and, according to technical regulations, should be referred to as *impregnante biocida* when containing the above-mentioned substances, while *Imprägnierung* always contain them. Besides, further distinction may be needed when referring to products that form a protective film on the wood surface. In this case, the most proper translation in German is *Lasur.* Nevertheless, *Lasur* is a hypernym that can identify two categories of products: *Dünnschichtlasur* on the one hand and *Dickschichtlasur* on the other hand. The most suitable target term can be identified only by means of an accurate analysis of the features and functions of the specific product referred to as *impregnante* and required close cooperation between the export department and German technicians. This cooperation allowed the employees to correct terminological mistakes and to implement the proper terminology in the software used internally by the company for the generation of multilingual technical documentation and in the terminological database managed externally by the translation agency.

Finally, the marketing and communications department oversees translations carried out by the translation agency and posts contents on the company's website. Newsletters, articles, and catalogues are sent to the translation agency, reviewed, and then published. In this case, language and terminology are not as specialised as in technical data sheets and must be clear and appealing to customers who visit the website. One of the most interesting findings concerned the English target term for the already-mentioned term *impregnante*. The most frequent target term used by employees and translators to translate it into English was *impregnating agent*. As opposed to the *Imprägnierung*, the issue does not consist of a partial equivalence nor of different practices among different countries but is in the actual use of the term abroad. Therefore, we carried out a two-fold analysis. First, we consulted websites of competitors to check the frequency of usage of the term *impregnating agent*. Then, we launched a search on the web and compared the frequency of *impregnating agent* with other candidate terms. We found out that the term *impregnating agent* as an English equivalent of the term *impregnante* is particularly widespread among Italian companies. On the contrary, English-speaking competitors usually refer to this category of products with the term *exterior wood stain*.

The following table shows the comparison carried out between Italian and English-speaking manufacturers, the number of websites consulted (second column), the number of websites where the term *impregnating agent* is used to designate the above-described product (third column) and other terms used with the same purpose by Italian companies as well as competitors (fourth column). As far as Italian companies are concerned, we focused the attention below on other incorrect terms used by them to designate the category of products called *impregnante,* in order to identify further examples of terminological mistakes and inaccuracies.

Companies	Number of websites	Occurrences of *impregnating agent*	Other terms used
Italian manufacturers of wood coatings[6]	44	18 websites out of 44	*Impregnating* *Filler* *Impregnating colour* *Impregnant* *Impregnation*
Foreign competitors[7] that produce wood coatings	20	None of them	*Exterior wood stain* *Exterior wood finish* *Wood preservative* *Deck and siding stain* *Finish for exterior wood*

Tab. 2 – *Comparison between English terms used by Italian companies and foreign competitors.*

We found that 18 out of 44 Italian companies use the term *impregnating agent* in the English version of their websites; 11 Italian companies out of 44 do not use the term *impregnating agent* but adopt other wrong terms such as *filler* – that is not equivalent to the term *impregnante* – or grammatically incorrect terms such as *impregnating* as a single-word term. Other terms, such as *impregnating colour*, *impregnant* and *impregnation* reveal further issues. The multi-word term *impregnating colour* combines the function of the product (*impregnating*) with the word *colour* that does not refer to a category of products but to a visual feature. Since a colour cannot be used with an impregnating function, the term seems to be quite vague. The terms *Impregnant* and *Impregnation* – like *impregnating agent* – are not widely used abroad and should be deprecated.

Analysis of 20 websites of wood coatings' manufacturers in USA, UK and Canada allowed us to test the above-mentioned hypothesis, namely the fact that

[6] We included in the analysis only Italian companies with an English version of their websites. This has however significantly reduced the number of websites consulted, since many Italian SMEs that sell wood coatings only have a website in Italian.

[7] The company mainly exports to USA, Canada, and UK. We therefore focused on these three English-speaking countries.

the term *impregnating agent* – frequently used as English target term at the observed company – is not employed by foreign competitors. On the contrary, they often use terms such as *exterior wood stain*, *exterior wood finish* and *wood preservative*. As far as American and British English are concerned, we did not identify significant differences in terms used in USA, Canada and UK. One regional variant is represented by the term *siding* (for instance in *deck and siding stain*), used mainly in the USA and in Canada but not in the UK, while other variants are due to different conceptual focuses rather than to regional practices. For example, in some cases competitors prefer to focus on the place of application of the product and use the modifier *exterior* (e.g., *exterior wood stain*), while in other cases they aim to highlight its protective function as in *wood preservative*. However, what stands out is the absence on competitors' websites of terms such as *impregnating agent*. The company has therefore adopted more competitive terms, such as *exterior wood stain* and *exterior wood finish*. These new terms have also been shared by the company with the translation agency and implemented in their translation memories and terminological database.

One last intervention on terminology used by the communications and marketing department has been carried out on the SEO list of keywords. SEO refers to strategies implemented by companies to increase user traffic on their websites. More specifically, website use of keywords that reflect the types of searches carried out by customers when looking for products online can have a positive impact on the ranking of the website in search engine results. As Warburton (2018) states: "it is therefore in a company's interest to align words on its website with those that are popular as search keywords for this type of content." Therefore, SEO keywords may be seen as a form of terminology since they must be relevant to the company's domain but also reflect the natural referencing processes of users. These keywords, selected on the basis of an accurate analysis of competitors' websites as well as through SEO tools, have their own peculiarities and must be shared with content producers and translators in order to guarantee their use in online content and translations. Working with SEO experts, the company already compiled a multilingual list of keywords related to their products, but this list had neither been updated nor revised by technicians and/or language experts. The following tables shows three examples of interventions that had to be carried out on the SEO list of keywords, along with some examples in English and German:

Type of intervention	Examples	Languages involved
Deprecation of wrong target terms	*Nagellack*	German
Addition of new terms related to new products produced by the company	*Wood sanitizer*	English
Addition of more competitive terms	*Clear wood finish* *Wetterschutzlasur*	English German
Substitution of existing terms with more competitive ones	*BIO coating → sustainable wood finish*, *eco-friendly wood finish* *BIO Lack → umweltfreundliche Lacke*, *Naturlack*	English German

TAB. 3 – SEO-related issues and improvement measures.

The first example shows the polysemic term *smalto* that refers to nail polish, as well as to a type of coating product for metal or wood. The equivalent term used in the German SEO set of keywords was not correct and was, therefore, deprecated.

The second example refers to a different type of intervention, i.e., the addition of a new term not previously included in the SEO set of keywords. During the pandemic, the company began producing coatings for wood sanitizing. Therefore, new keywords had to be implemented in order to attract customers looking for this specific product type. The term *wood sanitizer* could be considered a sort of "new entry" in the company-speak that – despite being well known outside of the company – has entered internal communication very recently and had to be supported by targeted SEO strategies as well as by consistent use on company communications channels.

The third example shows the addition of new terms related to already existing products. In this case, the aim was not to designate new products or processes but to widen the SEO strategy and account for keywords in widespread use abroad but not yet been included in the company's SEO list of terms. For instance, the adjective *clear* – as an attribute of *wood finish* – is widely used in the USA and was therefore included in the SEO strategy. Also, the above-mentioned term *Lasur* – as well as the multi-word term *Wetterschutzlasur* (i.e., protective exterior wood stain) – had not previously been in the company's SEO strategy, despite being the correct term to designate one of the main product types sold by the company.

Finally, some terms have been deprecated not because of being grammatically or conceptually incorrect, but because of very low engine search volume. In this case, terms were replaced with more competitive ones that better reflected the type of searches carried out by users in target countries. For instance, the multi-word term *BIO-coating* was replaced with *sustainable wood finish* and *eco-friendly wood finish*.

These keywords were then checked and approved by SEO experts and shared with the translation agency.

4. Conclusion

The above-mentioned issues showed the need for a company-wide analysis of terminology and cooperative work involving several experts. Intralingual terminological issues – such as improper use of terms and ambiguity arising from polysemic terms – can represent the initial stage of translation mistakes and interestingly reflect deviations in terms of specialised knowledge or specialised terminology from practices and standards established by the expert community. Interlingual terminological variants can be related not only to different cultures but also to manufacturing practices and industrial standards that differ from one country to another.

The collaborative work we carried out was deeply influenced by the type of translation workflows and texts involved. The export department translates technical data sheets, therefore the experts involved were technicians at the company's foreign branches. More specifically, we compared different candidate target terms as well as the concepts to which they refer in order to ensure that manufacturing practices and products were the same. As far as the communications and marketing department was concerned, cooperation with SEO experts and export managers was paramount to identify keywords that better reflected the actual usage of terms in foreign markets. In fact, when addressing foreign customers or when planning a SEO strategy, the terminology should reflect authentic terms used abroad by customers when searching a specific product on the Internet. The use of terms that are neither widespread nor known abroad may compromise the traffic on the company's website and have a negative impact on its competitiveness.

Finally, the direct involvement of different communities that deal on a day-to-day basis with terminology – but with different purposes and levels of specialised knowledge – allowed us to reach a two-fold objective: on the one hand, users were able to put a finger directly on issues related to wrong or inaccurate multilingual terminology and, on the other hand, we were able to improve the quality of multilingual terminology in a targeted way, according to the needs of each community of users. Nevertheless, only a constant update of the terminological database managed externally by the translation agency or even the implementation of an internal system will ensure the long-term quality of the terminology used. This second solution will, however, require more time and resources, both in terms of competences as well as technologies.

As a future research step, the aim is to extend the analysis to other local companies to identify and compare further terminological issues and verify whether an internal management of terminology could be adapted to different business environments.

References

Barsch-Harjau, Daniela. 2020. "Wo macht SEO-Terminologie im Unternehmen Sinn? Eine kleine Auseinandersetzung mit den Folgen der „SEO-Manie"." In *Terminologie: Industrie, Information, Intelligenz*, edited by Petra Drewer, Felix Mayer and Donatella Pulitano, 83-88. München/Karlsruhe/Bern: Deutscher Terminologie-Tag e.V.

Bertaccini, Franco & Lecci, Claudia. 2010. "La Variazione in Traduzione E in Redazione Tecnica: Verso Una Tipologia Differenziata Delle Schede Terminologiche." *Publifarum* 12

Delavigne, Valérie & de Vecchi, Dardo. 2016. "Socioterminologie et pragmaterminologie: rencontres et complémentarités." In *TOTh International Conference*, 141-156. Chambéry: Condillac.

de Vecchi, Dardo. 2012. "What do "they" mean by that? The (hidden) role of language in a merger." *LSP Journal* 3(2): 71-85.

de Vecchi, Dardo. 2016. "Approche pragmaterminologique des termes des enterprises et des organisations." *Synergies Italie* 12: 125-139.

de Vecchi, Dardo. 2020. "Words at Work: The Dynamics of Company-Speak in the Work Place." *Hermes – Journal of Language and Communication in Business* 60: 241-249.

Drewer, Petra. 2019. “Das Berufsbild „Terminologe/Terminologin“: Anforderungen, Qualifikationen, Ausbildung.” In *Terminologie: Epochen – Schwerpunkte – Umsetzungen*, edited by Petra Drewer and Donatella Pulitano, 59-81. Berlin/Heidelberg: Springer Vieweg.

Drewer, Petra & Schmitz, Klaus-Dirk. 2017. *Terminologiemanagement: Grundlagen - Methoden - Werkzeuge*. Berlin/Heidelberg: Springer Vieweg.

Kämer, Vivien. 2020. “Terminologiedatenbank mit Begriffssystem. Ein Terminologieverwaltungssystem für die Zukunft.” In *Terminologie: Industrie, Information, Intelligenz*, edited by Petra Drewer, Felix Mayer and Donatella Pulitano, 49-55. München/Karlsruhe/Bern: Deutscher Terminologie-Tag e.V.

Holdermann, Fabienne. 2020. “Terminologiemanagement in einem international agierenden Unternehmen von 0 auf 100 (gefühlt von 0 auf 100).” In *Terminologie: Industrie, Information, Intelligenz*, edited by Petra Drewer, Felix Mayer and Donatella Pulitano, 57-65. München/Karlsruhe/Bern: Deutscher Terminologie-Tag e.V.

ISO 1087-1. 2019. “Terminology work and terminology science – Vocabulary”. Geneva: International Standards Organisation.

ISO 4618. 2014. “Paints and varnishes – Terms and definitions”. Geneva: International Standards Organisation.

Lesk, Susanne, Lavric, Eva & Stegu, Martin. 2017. Multilingualism in business: Language needs.” In *Handbook of Business Communication*, edited by Gerlinde Mautner and Franz Rainer, 249-317. Boston/Berlin: De Gruyter Mouton.

Schmitz, Klaus-Dirk & Straub, Daniela. 2016. *Erfolgreiches Terminologiemanagement im Unternehmen: Praxishilfe und Leitfaden: Grundlagen, Umsetzung, Kosten-Nutzen-Analyse, Systemübersicht*. 2., aktual. Aufl. Stuttgart: tcworld.

Welch, Denice, Welch, Lawrence & Marschan-Piekkari, Rebecca. 2001. “The Persistent Impact of Language on Global Operations.” *Prometheus* 19(3), 193-209.

Résumé

Les entreprises d’envergure internationale doivent souvent traiter non seulement d’importants volumes de traduction, mais également les multiples terminologies linguistiques qui leur sont associées. Leur terminologie est profondément influencée par la pratique quotidienne de la langue, ainsi que par les communautés de personnes qui mobilisent cette terminologie, aussi bien au sein d'une même langue que dans les interactions d'une langue à une autre. Cette contribution vise à analyser les enjeux liés à la terminologie ainsi que leurs conséquences sur les flux de traduction dans le cadre d'une étude de cas spécifique, à savoir l'exemple d'une entreprise italienne de peinture pour le bois. A partir des conclusions préliminaires tirées d'une analyse quantitative menée dans la région italienne des Marches, nous procéderons à la présentation des problèmes liés à la terminologie tels qu'ils apparaissent dans le processus concret de traduction, notamment au sein d'une compagnie à échelle locale. L'objectif de cette étude est, d'une part, de présenter ces enjeux en prenant en considération chaque service impliqué dans le processus de traduction, et d'autre part, d'exposer l'analyse collaborative et interdisciplinaire appliquée afin de désambiguïser les variantes terminologiques et d'améliorer la qualité des traductions.

La gestion de (méta)données terminologiques «FAIR»: le répertoire de catégories de données de la ressource TriMED

Federica Vezzani

Département d'études linguistiques et littéraires (DiSLL)
Via Elisabetta Vendramini, 13, 35137, Padoue - Italie
Université de Padoue
federica.vezzani@unipd.it
http://www.dei.unipd.it/~vezzanif/

Résumé. Cette étude est consacrée à la description d'un flux de travail pour la gestion optimale de données et métadonnées terminologiques figurant dans les ressources terminologiques. Le paradigme de la «Terminologie FAIR» sur lequel nous sommes basées vise à fournir des lignes directrices pour la mise à disposition de données de la recherche terminologique trouvables, accessibles, interopérables et réutilisables. En particulier, nous nous concentrons sur la description d'un nouveau répertoire de catégories de données terminologiques - conçues comme des classes d'information - spécifiquement implémenté à l'appui de la ressource terminologique multilingue et polyvalente TriMED pour le domaine médical. Le répertoire est structuré selon les lignes directrices fournies dans la norme ISO 12620: 2019 et représente l'une des étapes vers la «FAIRification» des données terminologiques contenues dans TriMED.

1. Introduction

L'implémentation et la mise à disposition d'une base de données terminologique nécessitent de nombreux efforts de conception préliminaire afin d'organiser et de gérer de manière optimale les (méta)données terminologiques.

> «Gérer les données, c'est s'assurer que celles-ci sont correctement sélectionnées, décrites, préservées et rendues accessibles pour un traitement

> et/ou une réutilisation, et ce, bien au-delà du projet de recherche qui les a fait naître et les a exploitées au premier chef».

C'est ainsi que les auteurs Calderan et Millet (2015) définissent l'ensemble de tâches relevant de l'activité de «curation de données» (de l'anglais *data curation*). Cette notion identifie donc, généralement, l'ensemble de bonnes pratiques pour l'organisation optimale des données de recherche (Palmer *et al.* 2013), dont la responsabilité incombe inévitablement au chercheur qui produit ces données (McLure *et al.* 2014, Corti *et al.* 2019). L'ampleur et l'importance de la curation des données ont conduit à l'émergence d'une littérature et de normes détaillées régissant les actions requises (De Matos *et al.* 2004, Eaker 2016, Erkimbaev *et al.* 2019). À cet égard, le respect de ces exigences peut garantir non seulement la sécurité, mais également un enrichissement continu de la valeur qualitative des données scientifiques. À ce propos, un ensemble de lignes directrices a été publié par Wilkinson *et al.* (2016) dans le cadre de la plateforme européenne, librement accessible en ligne, *European Open Science Cloud* (EOSC)[1], afin de promouvoir la «FAIRness» des données de recherche. Ces lignes directrices soulignent la nécessité de données trouvables, accessibles, interopérables et réutilisables (*Findable, Accessible, Interoperable* et *Reusable*)[2]. Toutes les composantes du processus de recherche devraient bénéficier donc de l'application de ces principes, afin d'en garantir la transparence, la reproductibilité et la réutilisabilité.

La bonne gestion des données n'est pas un objectif en soi, mais plutôt le principal moyen menant à la découverte et à l'innovation des connaissances, ainsi qu'à l'intégration et à la réutilisation des données par la communauté scientifique. Dans le cadre de l'activité terminographique, l'Organisation internationale de normalisation (ISO)[3] et, en particulier, le Comité technique ISO/TC 37 (Langage et terminologie)[4] fournissent des normes spécifiques pour la conception et l'implémentation des ressources linguistiques et terminologiques structurellement homogènes. Toutefois, dans ce contexte, les données de la recherche sont encore loin d'être FAIR (Forkel *et al.* 2018). Les ressources langagières sont souvent encodées dans un format hétérogène et développées isolément les unes des autres (Cimiano *et al.* 2020), au risque de rendre leur découverte, leur réutilisation et leur intégration une tâche difficile et lourde. En ce sens, il faut noter les efforts menés par l'infrastructure

[1] https://www.eosc-portal.eu

[2] https://www.go-fair.org/fair-principles/

[3] https://www.iso.org/fr/home.html

[4] https://www.iso.org/fr/committee/48104/x/catalogue/

de recherche européenne CLARIN[5] qui permet aux chercheurs en sciences humaines d'accéder aux ressources et technologies linguistiques disponibles au niveau européen et vise à fournir une architecture des données conforme aux principes FAIR (De Jong *et al.* 2018).

Notre proposition s'inscrit dans ce contexte et vise à la description d'une étude de cas portant sur l'organisation FAIR des (méta)données terminologiques de la ressource multilingue TriMED[6] conçue pour le domaine d'application médical (Vezzani et Di Nunzio 2020b). La base de données terminologique a été implémentée dans l'esprit de la «terminologie FAIR» qui porte sur l'application des normes ISO les plus récentes sur la gestion des ressources terminologiques dans le but ultime de fournir à la communauté scientifique des données trouvables, accessibles, interopérables et réutilisables (Vezzani et Di Nunzio 2020a). En particulier, dans cette étude nous nous concentrerons sur le développement d'une ressource parallèle et complémentaire à TriMED, c'est-à-dire son répertoire de catégories de données[7] qui a été spécifiquement conçu, en respectant la norme ISO 12620 : 2019[8], afin de documenter, d'harmoniser et de gérer de manière optimale les (méta)données terminologiques. L'organisation de l'article est la suivante. La section 2 est consacrée à la description de la notion de «catégories de données terminologiques». La section 3 illustre le cadre général dans lequel s'inscrit le projet de recherche TriMED et définit le paradigme de la «Terminologie FAIR» pour la structuration optimale des ressources terminologiques. En particulier, nous nous concentrons sur la description du répertoire de catégories de données spécifiquement conçu pour la ressource TriMED. Enfin, dans la section 4, nous présentons nos conclusions et perspectives de recherche à long terme.

2. Catégories de données terminologiques

Les données figurant dans les ressources linguistiques sont généralisées dans des classes appelées «catégories de données». Une catégorie de donnée est une classe d'informations étroitement liées d'un point de vue formel ou sémantique et fait partie d'un schéma de collecte de données ou d'annotation pour une ressource linguistique donnée (ISO 12620 : 2019). Par exemple, /

[5] https://www.clarin.eu

[6] https://purl.org/trimed

[7] La ressource est disponible au lien suivant : https://shiny.dei.unipd.it/TriMED/data_category_repository/

[8] https://www.iso.org/fr/standard/69550.html ?browse=tc.

définition/ et /partie du discours/[9] sont des catégories de données communes dans les ressources terminographiques et lexicographiques.

Toutefois, les variétés d'approches pour développer différents types de ressources linguistiques entraînent inévitablement des variations dans les dénominations et les définitions de catégories de données. En ce sens, le concept de cohérence des catégories de données est fondamental et se produit dans leur dénomination et dans les valeurs que ces catégories admettent (Warburton et Wright 2019). À des fins d'interopérabilité, par exemple, imaginons que nous voulions intégrer automatiquement les données de deux ressources terminologiques différentes qui présentent toutes deux la catégorie de donnée / définition/. Si le nom de la catégorie est différent (par exemple *définition* d'une part et *déf.* d'autre part), il serait difficile pour un logiciel d'interpréter les deux noms comme figurant la même catégorie et donc leur intégration ne serait pas automatique, mais demanderait un travail chronophage de nettoyage manuel (ou semi-automatique) des données.

Afin d'assurer la cohérence entre les catégories de données terminologiques représentées dans différentes ressources terminologiques (et pour que celles-ci «dialoguent» du point de vue structurel), un répertoire normalisé de catégories de données a été mis à la disposition avec la publication de la dernière version de la norme ISO-12620 de 2019. Le répertoire TermWeb, disponible sur le site DatCatInfo[10], remplace l'ancienne ressource ISOcat développée et maintenue sous les auspices de l'ISO TC/37, avec l'Institut Max Planck de psycholinguistique de Nimègue, aux Pays-Bas, agissant en tant qu'autorité d'enregistrement (Kemps-Snijders *et al.* 2008, Broeder *et al.* 2014, Windhouwer et Schuurman 2014). Le nouveau répertoire TermWeb recueille une liste de toutes les catégories de données disponibles pour ceux qui souhaitent implémenter une ressource terminologique normalisée. Les catégories de données sont organisées en fiches descriptives et formelles (sous le nom de «spécifications») qui collectent toutes les informations nécessaires à leur documentation et à leur harmonisation, par exemple: nom, définition, exemples, commentaires, etc.

Par ailleurs, un aspect fondamental introduit par la norme ISO-12620 de 2019 est la possibilité pour tout implémenteur de concevoir son propre répertoire de catégories de donnée spécifique pour une ressource terminologique.

[9] À des fins de cohérence avec les normes, nous utiliserons les barres obliques inversées lorsqu'indiquant une catégorie de donnée.

[10] http://datcatinfo.net

Cette flexibilité de la norme est essentielle lorsqu'il est nécessaire de documenter des catégories de données qui n'existent pas dans le répertoire officiel TermWeb. De plus, les catégories de données ont des valeurs qui peuvent varier en fonction de la ressource conçue. La norme ISO-12620 de 2019 est complémentaire à la norme ISO 16642 : 2017[11] qui décrit le métamodèle structurel *Terminological Markup Framework* (TMF) à adopter pour la mise à disposition de ressources terminologiques interopérables. Le cadre de balisage terminologique TMF comprend un métamodèle structurel composé de trois entités principales et hiérarchiques (concept, langue et terme) et un ensemble de catégories de données pouvant être associées, selon une certaine liberté[12], à ces trois niveaux. En ce sens, la mise en œuvre parallèle d'un répertoire de catégories de données spécifiques pour une nouvelle ressource s'avère être une action indispensable non seulement pour la documentation correcte des (méta)données terminologiques, mais aussi pour désambigüiser leur signification et pour expliciter leurs contraintes en favorisant, de cette manière, le « dialogue structurel » entre les différentes ressources terminologiques et leur « FAIRness ».

3. Le projet TriMED

Avant de décrire l'objet de cette étude, à savoir l'élaboration d'un répertoire pour la documentation et la gestion optimale des catégories de données et des métadonnées terminologiques, nous présentons le cadre du projet de recherche dans lequel cette ressource s'inscrit. Le projet TriMED[13] découle de la volonté de fournir une ressource multilingue et polyvalente pour le domaine médical, et structurée afin de respecter les besoins sous-jacents aux pratiques de la science ouverte (Vezzani et Di Nunzio 2019 ; Vezzani et Di Nunzio 2020a). Dans ce contexte, le projet a envisagé le développement du paradigme de la « Terminologie FAIR » proposant des lignes directrices pour la conception et l'implémentation optimale des ressources terminologiques (Vezzani 2020). Le paradigme vise à définir un flux de travail pour la mise à la disposition de (méta) données terminologiques trouvables, accessibles, interopérables et réutilisables et est basé sur l'adoption des dernières normes ISO TC/37 SC/3 pour la gestion terminologique :

[11] https://www.iso.org/fr/standard/56063.html ?browse=tc

[12] Par exemple, selon la norme les éléments <term>, <termNoteGrp> et <termNote> peuvent uniquement apparaître au niveau du terme (<termSec>).

[13] https://purl.org/trimed

- la norme ISO16642 : 2017[14] qui définit le métamodèle structurel abstrait *Terminological Markup Framework* (TMF) pour la représentation des ressources terminologiques ;
- la norme ISO12620 : 2019[15] qui définit les propriétés des catégories de données et leur documentation dans un répertoire ; et
- la norme ISO30042 : 2019[16] qui définit le format de représentation *TermBase eXchange* (TBX) des données terminologiques.

Sur la base des lignes directrices du paradigme proposé, afin qu'une ressource terminologique puisse être FAIR, elle devrait être conçue et implémentée en :

- suivant un modèle structurel interopérable (TMF) ;
- permettant l'accès aux données terminologiques via des protocoles de communication standard ;
- fournissant des (méta)données rigoureusement documentées, et par conséquent trouvables, à travers un répertoire de catégories de données ;
- garantissant la réutilisation des données grâce à l'application de formats pour l'échange terminologique (TBX).

Ces principes ont été adoptés pour le développement de la base de données multilingue (italien, français et anglais) TriMED qui vise à satisfaire les besoins d'information de différentes catégories d'utilisateurs (patients, traducteurs et médecins) en fournissant un modèle de fiche terminologique contenant jusqu'à 42 catégories de données terminologiques. En particulier, la ressource a été conçue afin de :

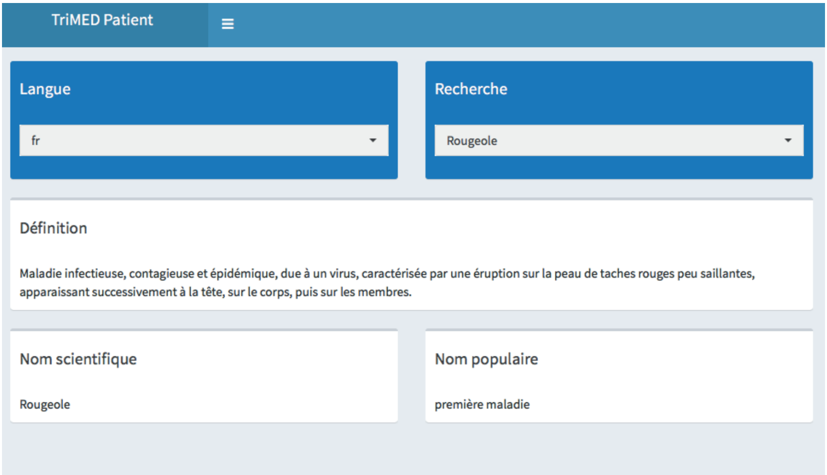
- aider les patients à comprendre correctement les informations médicales, compte tenu de l'aspect de variation diastratique de la terminologie ;
- soutenir le traducteur dans le processus de traduction spécialisée en fournissant un cadre sur le comportement syntaxique, sémantique et phraséologique du terme source et de son traduisant en langue cible ;
- fournir un point d'accès unique pour la consultation des professionnels de la santé aux autres terminologies, nomenclatures ou codes de classification internationaux généralement utilisés par des experts.

[14] https://www.iso.org/fr/standard/56063.html?browse=tc.

[15] https://www.iso.org/fr/standard/69550.html?browse=tc.

[16] https://www.iso.org/fr/standard/62510.html?browse=tc.

La figure 1 représente un exemple d'affichage de la ressource pour la catégorie d'utilisateurs « Patient ».



Langue

fr

Recherche

Rougeole

Définition

Maladie infectieuse, contagieuse et épidémique, due à un virus, caractérisée par une éruption sur la peau de taches rouges peu saillantes, apparaissant successivement à la tête, sur le corps, puis sur les membres.

Nom scientifique

Rougeole

Nom populaire

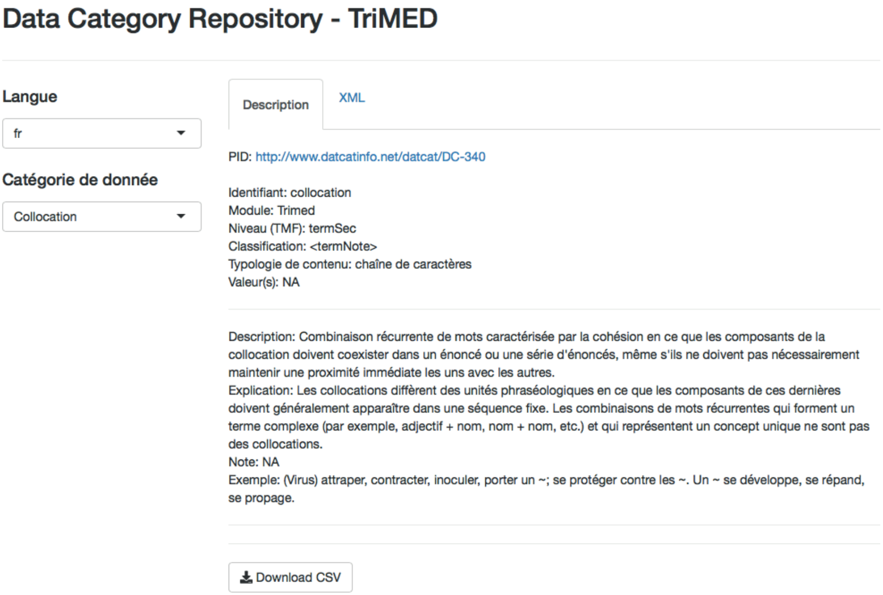
première maladie

Fig. 1 – *Affichage « Patient » - TriMED*

3.1. Répertoire de la ressource TriMED

Les exigences d'une structuration FAIR de la ressource TriMED nous ont menées à développer, en suivant la norme ISO-12620 de 2019, un répertoire parallèle et complémentaire contenant les spécifications de toutes les catégories de données affichées pour les utilisateurs. À notre connaissance, il s'agit du premier répertoire implémenté selon cette norme, dont l'application Web (déjà disponible en ligne)[17] a été développée à l'aide du package *Shiny R* (Winston *et al.* 2018). L'utilisateur peut tout d'abord sélectionner la langue de travail et, ensuite, saisir la dénomination de la catégorie dans la boîte de recherche : le système filtrera automatiquement les mots et les caractères indiquant les options possibles. Une fois sélectionnée la catégorie de donnée, le système affichera en sortie automatique la spécification de la catégorie contant toutes les informations requises par la norme afin d'en garantir la trouvabilité, l'accessibilité, l'interopérabilité et la réutilisabilité.

[17] https://shiny.dei.unipd.it/TriMED/data_category_repository/



Data Category Repository - TriMED

Langue

fr

Catégorie de donnée

Collocation

PID: http://www.datcatinfo.net/datcat/DC-340

Identifiant: collocation
Module: Trimed
Niveau (TMF): termSec
Classification: <termNote>
Typologie de contenu: chaîne de caractères
Valeur(s): NA

Description: Combinaison récurrente de mots caractérisée par la cohésion en ce que les composants de la collocation doivent coexister dans un énoncé ou une série d'énoncés, même s'ils ne doivent pas nécessairement maintenir une proximité immédiate les uns avec les autres.
Explication: Les collocations diffèrent des unités phraséologiques en ce que les composants de ces dernières doivent généralement apparaître dans une séquence fixe. Les combinaisons de mots récurrentes qui forment un terme complexe (par exemple, adjectif + nom, nom + nom, etc.) et qui représentent un concept unique ne sont pas des collocations.
Note: NA
Exemple: (Virus) attraper, contracter, inoculer, porter un ~; se protéger contre les ~. Un ~ se développe, se répand, se propage.

FIG. 2 – *Spécification de la catégorie de donnée /collocation/.*

La figure 2 ci-dessus montre une capture d'écran de la spécification de catégorie de donnée /collocation/. Pour chaque catégorie de données, les informations suivantes sont fournies:

1. Un identifiant unique et persistant (PID), c'est-à-dire une URL qui fournit l'accès Web direct à la spécification de la catégorie de donnée dans le répertoire en ligne.
2. Un identifiant mnémonique unique et stable de la catégorie de donnée qui ne doit pas inclure d'espaces entre les mots, car il est utilisé dans les environnements de codage comme élément ou comme valeur d'attribut.
3. Le module de catégories de données TBX auquel la catégorie se réfère[18].

[18] Le format d'implémentation choisi pour la ressource TriMED est le *TermBase Exchange* (TBX) conformément à la norme ISO 30042: 2019: https://www.iso.org/fr/stan-

4. Le niveau du métamodèle TMF (concept, langue et terme) auquel la catégorie de donnée est associée dans notre ressource.
5. La typologie de contenu de la catégorie de donnée, c'est-à-dire les types d'informations que la catégorie de donnée permet par sa mise en œuvre, comme une « liste déroulante » ou une « chaîne de caractères ».
6. L'ensemble de valeurs énumérées que la catégorie de donnée peut avoir si elle est implémentée en tant que « liste déroulante ».
7. La définition de la catégorie de donnée.
8. D'autres explications et notes sur la catégorie de donnée.
9. Quelques exemples d'utilisation de la catégorie de donnée.
10. La traduction du nom canonique de la catégorie de donnée dans les autres langues de travail de la ressource.

Il faut noter que, afin d'assurer une certaine traçabilité, les informations fournies dans le répertoire de TriMED concernant les PID des catégories de données se réfèrent à l'URL correspondant sur le site DatCatInfo pour les catégories qui sont déjà documentées dans le répertoire TermWeb (voir figure 2). Il existe également un nombre restreint de catégories de données qui sont exclusives pour notre ressource et ne sont donc pas illustrées sur DatCatInfo. Ces catégories de données terminologiques comprennent : 1) /analyse sémique/, 2) /hyperonyme/, 3) /hyponyme/, 4) /sous-domaine/, 5) /code ICPC2/, 6) /code ICD10/, 7) /terme SNOMED CT/, 8) /terme MeSh/ et 9) /sphère conceptuelle/. Pour ces catégories, nous fournissons un PID qui correspond à l'URL de la page associée dans notre répertoire TriMED. Par exemple, pour la catégorie de donnée /analyse sémique/ le PID fourni est : http://purl.org/trimed/dcr/dc/dc-1. En outre, comme suggéré par la norme et pour répondre au besoin de réutilisation des données, l'utilisateur peut exporter les informations fournies dans le répertoire dans le format Comma-Separated Values (CSV) et eXtensible Markup Language (XML). Enfin, le répertoire de TriMED diffère du répertoire TermWeb en ce que, contrairement à ce dernier dans lequel seules des traductions du nom canonique des catégories de données sont proposées, notre application fournit à l'utilisateur un système d'affichage multilingue complet. En ce sens, des informations telles que « description », « explication »,

dard/62510.html?browse=tc. Chaque catégorie de données appartient donc à un module TBX (public ou privé). Pour plus d'informations sur cet aspect, voir Vezzani et Di Nunzio (2020a).

« notes » et « exemples » sont adaptées en fonction de la langue de consultation de l'utilisateur. Ce type d'affichage multilingue se reflète également dans la génération automatique de fichiers d'exportation afin que l'utilisateur puisse télécharger et réutiliser une spécification dans un format lisible par la machine et adapté à sa langue de travail.

4. Conclusions et perspectives

La gestion FAIR des (méta)données de la recherche terminologique demande un énorme travail de conception préliminaire afin d'organiser des informations structurellement homogènes. Ce type d'approche nécessite également une certaine réflexion sur l'activité terminographique : le chercheur se trouve donc à se pencher non seulement sur l'élaboration des données terminologiques pertinentes pour le domaine d'étude, mais aussi sur leur structuration optimale afin de permettre leur partage et leur réutilisation par la communauté scientifique. Dans cette étude, nous avons décrit le développement du répertoire de catégories de données de TriMED, en tant qu'une première étape vers une « terminologie FAIR ». Les étapes présentées dans cet article couvrent la plupart des éléments clés du processus de FAIRification des données terminologiques. En perspective, nous entendons également définir les moyens de transformation de notre modèle sémantique en conformité avec le mouvement *Linguistic Linked Open Data* (LLOD). En ce sens, notre proposition contient un modèle conceptuel - schéma entité-association (Chen 1976) - qui peut être utilisé pour obtenir le graphe associé et transformer les données obtenues à partir de XML en RDF (*Resource Description Framework*) en tant que modèle de référence de données liées (Chiarcos *et al.* 2012).

Enfin, nous travaillons également à l'adoption de la même méthodologie d'organisation des données dans le cadre du projet européen « Terminologie sans frontières » promu par l'Unité de Coordination de la Terminologie (TermCoord) du Parlement européen. Le projet implique la collaboration de nombreux partenaires et requiert une action nécessaire de documentation et d'harmonisation des données dans l'esprit de la science ouverte. Dans ce cadre, nous avons mis en place et testons l'ergonomie de la nouvelle application web « FAIRterm[19] » (Vezzani 2021) comme outil de compilation de fiches terminologiques multilingues dans le cadre du projet européen. Cet outil vise donc à mettre à disposition des terminologues travaillant dans toute l'Europe une plateforme en accès libre permettant l'échange et le partage de données.

[19] https://purl.org/fairterm.

Références

Broeder, Daan, Ineke Schuurman, and Menzo Windhouwer. 2014. “Experiences with the ISOcat data category registry.” In *LREC 2014 : 9th International Conference on Language Resources and Evaluation*, pp. 4565-4568.

Calderan, L. et J. Millet. 2015. B*IG DATA : nouvelles partitions de l’information : Actes du séminaire IST Inria, octobre 2014*. Louvain-la-Neuve : De Boeck Superieur.

Winston, C., J. Cheng, J. J. Allaire, Y. Xie, et J. McPherson. 2018. “Shiny : Web Application Framework for R. R package version 1.1. 0.”

Chen, Peter Pin-Shan. 1976. “The entity-relationship model—toward a unified view of data.” *ACM transactions on database systems (TODS)* 1, no. 1 : 9-36.

Chiarcos, Christian, Sebastian Nordhoff, et Sebastian Hellmann. 2012. *Linked Data in Linguistics*. Heidelberg : Springer.

Cimiano, Philipp, Christian Chiarcos, John P. McCrae, et Jorge Gracia. 2020. *Linguistic Linked Data*. Cham, Switzerland : Springer International Publishing.

Corti, Louise, Veerle Van den Eynden, Libby Bishop, et Matthew Woollard. 2019. *Managing and sharing research data : a guide to good practice*. Los Angeles : SAGE Publications Limited.

De Matos, David Martins, Ricardo Daniel Santos Faro Marques Ribeiro, et Nuno J. Mamede. 2004. “Rethinking reusable resources”. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC’04), Lisbon, Portugal. European Language Resources Association (ELRA).

De Jong, F. M. G., Bente Maegaard, Koenraad De Smedt, Darja Fišer, et Dieter Van Uytvanck. 2018. “CLARIN : towards FAIR and responsible data science using language resources.” In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 3259-3264.

Eaker, Chris. 2016. “What could possibly go wrong ? the impact of poor data management”. In *The Medical Library Association’s Guide to Data Management for Librarians*. Lanham, Maryland : Rowman and Littlefield Publishing Group.

Erkimbaev, A. O., V. Y. Zitserman, G. A. Kobzev, et A. V. Kosinov. 2019. “Curation of digital scientific data”. *Scientific and Technical Information Processing* 46 (3), 192-203.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, et Russell D. Gray. 2018. "Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics." *Scientific Data* 5, no. 1 : 1-10.

ISO-12620 (2019). *Management of terminology resources – Data category specifications*. Standard, International Organization for Standardization, Geneva, CH.

ISO-16642 (2017). *Computer applications in terminology – Terminological markup framework*. Standard, International Organization for Standardization, Geneva, CH.

ISO-30042 (2019). *Management of terminology resources – TermBase eXchange (TBX)*. Standard, International Organization for Standardization, Geneva, CH.

Kemps-Snijders, Marc, Menzo Windhouwer, Peter Wittenburg, et Sue Ellen Wright. 2008. "ISOcat: Corralling data categories in the wild." In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.

McLure, Merinda, Allison V. Level, Catherine L. Cranston, Beth Oehlerts, et Mike Culbertson. 2014. "Data curation: a study of researcher practices and needs." *portal: Libraries and the Academy* 14, no. 2: 139-164.

Palmer, Carole, Nicholas M. Weber, Allen H. Renear, et Trevor Muñoz. 2013. "Foundations of data curation: The pedagogy and practice of "purposeful work" with research data". *Archives Journal* 3.

Vezzani, Federica, et Giorgio Maria Di Nunzio. 2019. "Computational terminology in eHealth". In *Italian Research Conference on Digital Libraries*, pp. 72-85. Springer.

Vezzani, Federica. 2020. *Vers une "Terminologie FAIR": conception et implémentation de la ressource TriMED,* Thèse de doctorat, Université de Padoue.

Vezzani, Federica et Giorgio Maria Di Nunzio. 2020a. "Methodology for the standardization of terminological resources: design of TriMED database to support multi-register medical communication". *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 26 (2), 266-298.

Vezzani, Federica et Giorgio Maria Di Nunzio. 2020b. "On the Formal Standardization of Terminology Resources: The Case Study of TriMED." In *Proceedings of the 12th Language Resources and Evaluation*

Conference, Marseille, France, pp. 4903-4910. European Language Resources Association.

Vezzani, Federica. 2021. “La ressource FAIRterm : entre pratique pédagogique et professionnalisation en traduction spécialisée”, *Synergies Italie*, n. 17., p. 51-64.

Warburton, Kara, et Sue Ellen Wright. 2019. “A data category repository for language resources.” *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, 69.

Wilkinson, Mark D., Michel Dumontier, Ij J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg *et al*. 2016. “The FAIR Guiding Principles for scientific data management and stewardship.” *Scientific data*, *3*(1), 1-9.

Windhouwer, Menzo, et Ineke Schuurman. 2014. “Linguistic resources and cats : how to use ISOcat, RELcat and SCHEMAcat.” In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pp. 3806-3810. European Language Resources Association (ELRA).

Abstract

This study aims at the description of a workflow for the optimal management of terminological data and metadata collected within terminological resources. We focus on the “FAIR Terminology” paradigm aiming to provide guidelines for the provision of findable, accessible, interoperable, and reusable terminology research data. In particular, we focus on the description of a new Repository of Data Categories - designed as information classes - specially implemented in support of the multilingual and multipurpose terminological resource TriMED for the medical domain. The repository is structured according to the guidelines provided in ISO 12620 : 2019 and represents one of the steps towards the “FAIRification” of the terminological data contained in TriMED.