

## **Terminological Resources in the Digital Age**

Christophe Roche, Amparo Alcina & Rute Costa

In a globalised society, terminological dictionaries – including resources such as knowledge and terminological databases, ontologies, wordnets, “traditional” dictionaries, etc. – should comply with both human and machine needs. Changes regarding information and language processing brought forward by the evolution of society have led to a series of consequences in: (i) the design of terminological resources; (ii) the way data and knowledge are represented; (iii) the way data are interrelated, both within and between resources; (iv) the way users access data; and (v) users’ expectations.

Moreover, in recent years, the creation of a number of electronic terminological resources has involved theoretical, methodological and technical advances, which have provided solutions to different issues related to content, search techniques and resource maintenance. At present, a large number of terminological resources are available in different formats and in different languages. We believe that now, after a number of years working on these topics, it is time to explore this new reality.

This special issue, consisting of 5 articles, does not claim to cover all aspects of terminological dictionaries. A single special issue would not be enough.

We have therefore decided to focus on the use and integration of terminological resources for the construction of electronic dictionaries in the digital age, particularly in the linked and open data and the semantic web context. Open data – and more precisely open access to linguistic resources such as Wordnet, terminological resources such as IATE (InterActive Terminology for Europe), encyclopaedic resources such as Wikipedia (DBpedia) and BabelNet, and highly specialised

Roche, Christophe, Amparo Alcina y Rute Costa. 2019. "Terminological Resources in the digital age", *Terminology*, 25 (2), 139-145. DOI: <https://doi.org/10.1075/term.00033.roc>.

Enlace de la editorial: <https://www.jbe-platform.com/content/journals/10.1075/term.00033.roc>

terminological resources such as Unified Medical Language System (UMLS) – allow us to have a different perspective on how new terminological resources and terminological dictionaries are built. The extraction of candidate terms, definition patterns and contexts, but also the processing of parallel corpora for translation purposes that is still one of the main applications of terminology today, are some examples. But this openness also imposes its own constraints. Data must be interoperable (APIs) and must be stored in a compatible exchange format, where the choice between the different available standards is not always easy, whether they come from standard organisations such as International Organization for Standardization (ISO) (e.g., TermBase eXchange - TBX) and Lexical Markup Framework - LMF) or the Semantic Web (OntoLex-Lemon for the linguistic dimension, Simple Knowledge Organization System (SKOS) or Ontology Web Language (OWL) for the conceptual dimension). This openness also impacts the structure and design of terminological databases, sometimes requiring their architecture and data structure to be reconsidered. It also requires appropriate construction methodologies. Finally, it is not enough for terminological resources to be interoperable and open to be fully exploitable. For example, how can we combine terminological resources in the same domain when their conceptual systems do not match?

The first article entitled "IATE 2: Modernising the EU's IATE terminological database to respond to the challenges of today's translation world and beyond" and proseed by Paula Zorrilla-Agut and Thierry Fontenelle (Translation Centre for the Bodies of the European Union) perfectly illustrates this approach. IATE (InterActive Terminology for Europe) is one of the most important terminological resources, probably the largest multilingual terminological database in the world. It covers all fields of activity of the EU and mostly contains terminology in the 24 official EU languages, although it also provides some content in non-official languages. This article presents the modernisation of the EU's IATE terminological database to respond to the challenges of today's translation world, including interoperability, modularity, scalability, and data exchange. It describes

Roche, Christophe, Amparo Alcina y Rute Costa. 2019. "Terminological Resources in the digital age", *Terminology*, 25 (2), 139-145. DOI: <https://doi.org/10.1075/term.00033.roc>.

Enlace de la editorial: <https://www.jbe-platform.com/content/journals/10.1075/term.00033.roc>

which strategies are being put in place to align IATE with the latest software and systems engineering standards and technologies, as well as to allow IATE data to be consumed by third-party tools, particularly computer-assisted translation environments. The new version of IATE has been conceived as a fully interoperable system with a responsive design built on a web service layer providing APIs. The IATE data structure is widely based on the data categories of the ISO 12620 standard, and many mechanisms have been implemented to manage the excessive volume of data and ensure their quality.

Creating dictionaries requires linguistic resources on which the quality and the richness of the dictionaries have to rely. Today, more and more linguistic resources are available in open access, even if that access remains limited mainly due to their dispersion and heterogeneous formats. Combined with natural language processing, they constitute an important and essential source of information to create e-dictionaries and terminological dictionaries providing useful information to be included in dictionary entries (e.g., offering help for users to understand the terms to be translated in texts). Using and integrating open resources in terminological e-dictionaries require new methodologies to produce such dictionaries, which will depend on the target application and available resources, for example, parallel corpora can be used for translation purposes. Furthermore, the heterogeneousness of terminological resources raises the problem of aligning their conceptual systems.

An example of such a methodology is provided by the second article entitled "Using open data to create the Catalan IATE e-dictionary". Written by Mercè Vázquez and Antoni Oliver (Universitat Oberta de Catalunya), and Elisabeth Casademont (TERMCAT), it aims to describe how the Catalan IATE e-dictionary is being produced given that IATE does not yet include Catalan. The method uses two types of resources: terminological dictionaries and parallel corpora. Terminological dictionaries come from the TERMCAT Terminologia Oberta terminology sets for the subject

Roche, Christophe, Amparo Alcina y Rute Costa. 2019. "Terminological Resources in the digital age", *Terminology*, 25 (2), 139-145. DOI: <https://doi.org/10.1075/term.00033.roc>.

Enlace de la editorial: <https://www.jbe-platform.com/content/journals/10.1075/term.00033.roc>

domains of economics, law and health, as well as two Terminologia Oberta terminology compilations containing terms from different subject domains. The terms from both IATE, which can be downloaded as a TBX (TermBase eXchange) file, and TERMCAT are organised by subject domain. We should highlight that the matching between the subject domains of the two resources has been manually performed. Wikipedia was also used as a terminological resource, albeit its categories do not always match IATE's domains. Since the main goal of this work is to produce the Catalan version of the IATE database, a parallel Spanish-Catalan corpus of texts was used to identify translation equivalents in IATE corresponding to the subject domains of law and administration. This corpus is extracted from the *Diari Oficial de la Generalitat de Catalunya*, an official media in which the laws and regulations of the Government of Catalonia are published in both languages, Spanish and Catalan.

The design of dictionaries must take into account multiple parameters: type of dictionary (general or specialised), type of dictionary users (professional users, non-professional users, language learners), and type of dictionary uses (learning, translating, writing specialised texts), as well as the characteristics of the subject itself. Also, in the case of multilingual dictionaries for translation purposes, there is the problem of comparing not only languages but also the realities of which these languages speak.

The third article is entitled "eLex – a Prototype Electronic Dictionary Application for Legal Translators" and it was written by Weronika Szemińska (University of Warsaw) and Adrian Więch. It is about designing dictionary software for translation in specific domains. Based on an analysis of target user needs and the characteristics of the subject field in question, the authors show that it is possible to specify a series of parameters that should be met by the final product. This approach is illustrated with eLex2, a prototype electronic dictionary application for legal translators. The idea is to divide the dictionary into five sub-dictionaries. The first four dictionaries (Explicative Dictionary,

Roche, Christophe, Amparo Alcina y Rute Costa. 2019. "Terminological Resources in the digital age", *Terminology*, 25 (2), 139-145. DOI: <https://doi.org/10.1075/term.00033.roc>.

Enlace de la editorial: <https://www.jbe-platform.com/content/journals/10.1075/term.00033.roc>

Contrastive Dictionary, Dictionary of Equivalence, and Combinatorial Dictionary) correspond to the first four stages of the translation process: source text reception, legal comparison, transfer, and target text production, respectively. The fifth sub-dictionary, the Concise Dictionary, closes the system.

Terminology has a double dimension – a linguistic dimension and a conceptual one. The ISO 1087 Standard recalls that a term is a “designation that represents a general concept by linguistic means” (ISO/FDIS 1087: 2019). However, while everyone cites the concept as being at the heart of terminological work, few systems define the concept theory on which they are based. And even fewer systems indicate how the conceptual system is represented, i.e., the formal language in which concepts are defined, when they are actually defined, which is not always the case. For some terminological systems, a concept is reduced to the term that designates it, or to a set of words referring to it, e.g., the Synsets of WordNet, and the conceptual system is reduced to a simple hierarchy of terms. Expressing domain knowledge in a formal language has many advantages. Concepts are extra-linguistic knowledge, this means they are independent of any natural language, even if this statement is certainly too strong: “Concepts are not necessarily bound to particular natural languages. They are, however, influenced by the social or cultural background which often leads to different categorizations” (Note 1 to entry 3.2.7 Concept, ISO/FDIS 1087: 2019). Based on the conceptual structure, it is possible to define new terminological system architectures, thus providing new features: interactive and graphical browsing guided by conceptual maps, and integrating and linking open terminological resources are some examples.

The fourth article, “EcoLexicon and by-products: integrating and reusing terminological resources” written by Pilar León-Araúz, Arianne Reimerink and Pamela Faber (University of Granada), is an illustration of this approach relying on the explicit representation of domain knowledge. This article addresses the problem of reusing and integrating terminological resources

Roche, Christophe, Amparo Alcina y Rute Costa. 2019. "Terminological Resources in the digital age", *Terminology*, 25 (2), 139-145. DOI: <https://doi.org/10.1075/term.00033.roc>.

Enlace de la editorial: <https://www.jbe-platform.com/content/journals/10.1075/term.00033.roc>

when building new applications. This is a major problem given the considerable time and effort involved in the development of terminologies. EcoLexicon is not only a multilingual terminological database but also a domain knowledge base that integrates conceptual, linguistic, and visual information - on environmental science. It relies on a theory of knowledge dedicated to non-language-specific representations, a theory that uses certain aspects of Frame Semantics (Fillmore). This underlying conceptual structure constitutes the backbone of the system. This approach allows writing definitions directly referring to the conceptual structure of the domain, as well as graphical navigation into resources based on a multi-level semantic network, while attaching multi-media information (e.g. images) to concepts, etc. Designed to help translators and technical writers, EcoLexicon gives access to a wide range of information by integrating, reusing and providing access to terminological resources – not only linguistic (corpus), but also terminological (IATE) and encyclopaedic (Babelnet, DPpedia) – as well as to dedicated software (Sketch Engine, Babelfy). This article details the EcoLexicon system and three of its by-products: the EcoLexicon English corpus, available as an open corpus in Sketch Engine; EcoLexiCAT, a terminology-enhanced translation tool providing access to domain-specific terminological knowledge in context; and Manzanilla, an image annotation tool to enhance knowledge representation in the Terminological Knowledge Base.

The last article offers a different perspective of the design and structure of terminological dictionaries, which was already outlined in the previous article. This perspective brings the conceptual dimension to the fore. This is made possible by the contribution of artificial intelligence, and in particular ontology from knowledge engineering and the semantic web. While the notion of concept is at the heart of any terminology, few approaches explicitly represent concepts. Expressing the concept as extra-linguistic knowledge in the form of a formal ontology (i.e., defined in a formal language understandable by a computer) makes it possible to agree on the only thing that can be

Roche, Christophe, Amparo Alcina y Rute Costa. 2019. "Terminological Resources in the digital age", *Terminology*, 25 (2), 139-145. DOI: <https://doi.org/10.1075/term.00033.roc>.

Enlace de la editorial: <https://www.jbe-platform.com/content/journals/10.1075/term.00033.roc>

standardised, namely domain knowledge, and to preserve what must be preserved, namely linguistic diversity. However, this raises several problems, such as the construction of the conceptual system or the relationships between terms and concepts that can no longer be confused because they fall within different semiotic systems, as well as the relationship between the formal definition of concepts and the definition of terms in natural language. While recognising the importance of the linguistic dimension of Terminology, and by clarifying the concept defined in a formal and specific language, this approach perceives the definition as a linguistic paraphrase of the logical definition of the concept denoted by the term.

Proposed by Christophe Roche (University Savoie Mont-Blanc), and Rute Costa, Sara Carvalho and Bruno Almeida (Universidade Nova de Lisboa), the last article, entitled "Knowledge-based terminological e-dictionaries: The EndoTerm and al-Andalus Pottery projects", illustrates this approach. It focuses on the double dimension of terminology and more precisely on modelling the conceptual system of a terminological dictionary. The authors show the full value of ontologies for conceptual system building and for the operationalisation of such dictionaries. After recalling the importance of the concept theory on which terminology is based, the authors present 3 different environments to construct conceptual systems. CmapTools is a graphical editor that is easy to use and suitable for illustrative purposes but lacks any consistency check. Protégé is an environment to build formal ontologies based on description logic. The Protégé ontologies are expressed in OWL (Ontology Web Language), the W3C language for ontologies. Finally, Tedi is a specific environment for the construction of ontoterminologies – terminologies whose conceptual system is a formal ontology, exportable in different formats including those of the W3C Standards (OWL, SKOS). The approach presented in this article is broad enough to be illustrated in fields as different as medicine and cultural heritage.

Roche, Christophe, Amparo Alcina y Rute Costa. 2019. "Terminological Resources in the digital age", *Terminology*, 25 (2), 139-145. DOI: <https://doi.org/10.1075/term.00033.roc>.

Enlace de la editorial: <https://www.jbe-platform.com/content/journals/10.1075/term.00033.roc>

In conclusion, the existence of many open resources, whether terminological or encyclopaedic, is today an essential source of information for the construction of terminological resources and terminological dictionaries. However, these often heterogeneous resources also raise many problems regarding format exchange, interoperability, integration, and linking. Hence, the architecture of terminological systems must be redesigned accordingly. For example, the integration of different terminological resources, even within the same domain, raises the problem of mapping their conceptual systems, a problem that is often underestimated if not outright ignored. The resources produced must also be interoperable or at least expressed in an exchange format (ISO, W3C). The conceptual system then appears as the backbone of these new systems, a conceptual system that natural language processing tools have so far pushed into the background. Such architectures offer new features, particularly when it comes to search and navigation capabilities within terminological resources, but also as to their enhancement, e.g., including images. Domain knowledge must be taken into account and represented in a formal language that computers are able to understand. In this context, the semantic web offers promising perspectives. At a "syntactic" level, we can mention the exchange formats for terminological resources. At a "semantic" level, let us quote the alignment of ontologies to map the conceptual systems of the terminologies to be integrated.

The scientific committee of this special issue is composed of the following members:

Guadalupe Aguado, Universidad Politécnica de Madrid, Spain

Bassey E. Antia, University of the Western Cape, South Africa

Gladis Barcellos, Universidade Federal de São Carlos, Brazil

Caroline Barrière, Computer Research Institute of Montréal, Canada

Nicoletta Calzolari, CNR-ILC, Italy



Roche, Christophe, Amparo Alcina y Rute Costa. 2019. "Terminological Resources in the digital age", *Terminology*, 25 (2), 139-145. DOI: <https://doi.org/10.1075/term.00033.roc>.

Enlace de la editorial: <https://www.jbe-platform.com/content/journals/10.1075/term.00033.roc>

Patrick Drouin, University of Montreal, Canada

Pamela Faber, Universidad de Granada, Spain

Purificación Fernández Nistal, Universidad de Valladolid, Spain

Thierry Fontenelle, Translation Centre for the Bodies of the European Union, Luxembourg

Rufus Gouws, University of Stellenbosch, South Africa

John Humbley, Université Paris Diderot, France

Ilan Kernerman, Kdictionaries, Israel

Patrick Leroyer, Aarhus University, Denmark

Teresa Lino, Universidade NOVA de Lisboa, Portugal

Pam Peters, Macquarie University, Australia

Carles Tebé, Pontificia Universidad Católica de Chile

Kara Warburton, Termologic, Hong Kong and Canada

#### Corresponding author

Christophe Roche  
Condillac Group – Listic Lab.  
Université Savoie Mont-Blanc  
Campus Scientifique  
73 376 Le Bourget du Lac cedex – France  
[christophe.roche@univ-savoie.fr](mailto:christophe.roche@univ-savoie.fr)

#### Co-authors information

Amparo Alcina  
*Tecnolettra Research Group*  
Translation and Communication Department  
Universitat Jaume I  
Campus del Riu Sec, s/n  
12071 Castellón -Spain  
[alcina@uji.es](mailto:alcina@uji.es)

Rute Costa  
NOVA – CLUNL, Centro de Linguística da Universidade NOVA de Lisboa

Texto original enviado para su publicación en:

Roche, Christophe, Amparo Alcina y Rute Costa. 2019. "Terminological Resources in the digital age", *Terminology*, 25 (2), 139-145. DOI: <https://doi.org/10.1075/term.00033.roc>.

Enlace de la editorial: <https://www.jbe-platform.com/content/journals/10.1075/term.00033.roc>

Avenida de Berna 26-C  
1069-061 Lisboa – Portugal  
[rute.costa@fcsh.unl.pt](mailto:rute.costa@fcsh.unl.pt)