

# Multilingual digital terminology: Introduction to the special issue

## 1 Introduction

Digital Terminology is a novel approach to the classical discipline of Terminology that includes all the activities for the design, implementation, and management of digital resources collecting terminological data and metadata for a specialized domain (Vezzani, 2022).

In this context, the terminologist of the digital age is not only concerned with the conceptual and linguistic representation of terminology, but also open to dialogue with the structural requirements necessary to preserve and manage a digital resource. Indeed, this new interdisciplinary figure should handle both the quality of the terminological data and the way the terminological resource is structured to ensure, for example, its accessibility and reuse.

The principles of findability, accessibility, interoperability, and reusability (FAIR) of digital resources (Wilkinson *et al.*, 2016) promoted by the European Open Science Cloud<sup>1</sup> Association have contributed significantly to shape the main ideas behind the Digital Terminology approach.

This scenario also paved the way for the foundation of the FAIR terminology paradigm.<sup>2</sup> FAIR terminology is the first initiative aiming at providing a set of guidelines for the optimal organization of terminological data, metadata, and infrastructure compliant with the FAIR principles (Vezzani and Di Nunzio, 2020; Vezzani, 2021, 2022). The paradigm is based on the adoption of the main ISO TC 37/SC 3 standards for the management of terminology resources.

## 1 Introduction

La terminologie numérique est une nouvelle approche de la discipline classique de la terminologie qui comprend toutes les activités de conception, d'implémentation et de gestion de ressources numériques collectant des données et des métadonnées terminologiques pour un domaine spécialisé (Vezzani, 2022).

Dans ce contexte, le terminologue de l'ère numérique est non seulement concerné par la représentation conceptuelle et linguistique de la terminologie, mais également ouvert au dialogue avec les exigences structurelles nécessaires à la préservation et à la gestion d'une ressource numérique. En effet, cette nouvelle figure interdisciplinaire devrait s'occuper à la fois de la qualité des données terminologiques et de la structuration de la ressource terminologique pour assurer, par exemple, son accessibilité et sa réutilisation.

Les principes de *Findability*, *Accessibility*, *Interoperability*, et *Reusability* (FAIR) des ressources numériques (Wilkinson *et al.*, 2016) promus par l'Association *European Open Science Cloud* (EOSC)<sup>1</sup> ont contribué de manière significative à façonner les idées principales derrière l'approche de la terminologie numérique.

Ce scénario a également ouvert la voie à la fondation du paradigme de la terminologie FAIR.<sup>2</sup> La terminologie FAIR est la première initiative visant à fournir un ensemble de lignes directrices pour une organisation optimale des données, des métadonnées et des infrastructures terminologiques conformes aux principes FAIR (Vezzani et Di Nunzio, 2020; Vezzani, 2021, 2022). Le paradigme est basé sur l'adoption des principales normes ISO TC 37/SC 3 pour la gestion des ressources terminologiques.

This is also the rationale behind the choice of the topics of this issue that explicitly refer to the design, representation formats, and implementation of multilingual terminological management systems.

This Special Issue of *Digital Scholarship in the Humanities* arises from discussions held at the first International Conference on ‘Multilingual digital terminology today. Design, representation formats and management systems’, MDTT 2022 (Di Nunzio *et al.*, 2022) held in Padova (Italy) on the 16 and 17 June 2022.<sup>3</sup>

The original idea behind the organization of a brand-new conference on Multilingual Digital Terminology, now at its second edition,<sup>4</sup> stemmed from the desire to give voice to all the interdisciplinary components of this active research community and to create a forum for the discussion, sharing of innovative ideas, and foster an open collaboration among researchers.

To promote a wider participation and dissemination of the results, this Special Issue (as well as the conference) is bilingual: the collected papers can be either in English or in French (with a bilingual abstract) and this introduction is bilingual.

## 2 Content of this special issue

The eight papers appearing in the current issue were selected among eleven submissions after an invited call for papers for the eighteen contributions to MDTT 2022. These selected papers stand out for their quality and relevance to the topic of Multilingual Digital Terminology.

On the one hand, the multilingualism component is addressed through the study and analysis of a total of six languages (Italian, English, French, Portuguese, Spanish, and Greek). On the other hand, the multidisciplinary aspect is tackled by the different subject fields, such as: natural language processing, environment, medicine, cultural heritage.

The heterogeneous research topics covered by this Special Issue range from knowledge organization, data modelling and standardization to corpus linguistics, automatic term extraction (ATE) and machine translation, including user needs and user studies.

The paper by Melania Cabezas-García and Pilar León-Araúz (‘Machine versus corpus-based translation of multiword terms’) analyses the implications of the work of a terminographer in the context of machine translation post-editing. With a view to enhancing the multilingual representation of environmental multiword terms in terminological resources, the authors analyse English–Spanish multiword term translation in various generic machine translation systems.

C’est aussi la raison d’être du choix des sujets de ce numéro spécial qui font explicitement référence à la conception, aux formats de représentation et à la mise en œuvre de systèmes de gestion terminologique multilingues.

Ce numéro spécial de *Digital Scholarship in the Humanities* découle des discussions tenues lors du 1er colloque international sur « La terminologie numérique multilingue aujourd’hui. Conception, formats de représentation et systèmes de gestion », MDTT 2022 (Di Nunzio *et al.*, 2022) qui s’est tenu à Padoue (Italie) les 16 et 17 juin 2022.<sup>3</sup>

L’idée originale de l’organisation d’un tout nouveau colloque sur la terminologie numérique multilingue, déjà à sa deuxième édition,<sup>4</sup> est née de la volonté de donner la parole à toutes les composantes interdisciplinaires de cette active communauté de recherche et de créer un lieu d’échange, de partage des idées novatrices et de favoriser une collaboration ouverte entre les chercheurs.

Pour encourager une plus large participation et diffusion des résultats, ce numéro spécial (ainsi que le colloque) est bilingue: les articles collectés peuvent être soit en anglais, soit en français (avec un résumé bilingue) et cette introduction est bilingue.

## 2 Contenu de ce numéro spécial

Les 8 articles apparaissant dans le présent numéro ont été sélectionnés parmi 11 soumissions après un appel à contributions sous invitation pour les 18 contributions à MDTT 2022. Ces articles sélectionnés se distinguent par leur qualité et leur pertinence par rapport au sujet de la terminologie numérique multilingue.

D’une part, la composante du multilinguisme est abordée à travers l’étude et l’analyse d’un total de 6 langues (italien, anglais, français, portugais, espagnol, grec). D’autre part, l’aspect pluridisciplinaire est abordé par les différents champs thématiques, tels que: traitement automatique du langage naturel, environnement, médecine, patrimoine culturel.

Les sujets de recherche hétérogènes couverts par ce numéro spécial vont de l’organisation des connaissances, la modélisation et la normalisation des données à la linguistique de corpus, l’extraction automatique de termes et la traduction automatique, en passant par les besoins des utilisateurs et les études d’utilisateurs.

L’article de Melania Cabezas-García et Pilar León-Araúz (*Machine versus corpus-based translation of multiword terms*) analyse les implications du travail d’un terminographe dans le contexte de la post-édition de la traduction automatique. Dans le but d’améliorer la représentation multilingue des termes environnementaux dans les ressources terminologiques, les auteurs analysent la traduction de termes composés anglais-espagnol dans divers systèmes génériques de traduction automatique.

The work by Rute Costa, Ana Salgado, Margarida Ramos, Sara Carvalho, Anas Khan, Toma Tasovac, Bruno Almeida, Mohamed Khemakhem, Laurent Romary, and Raquel Silva ('A crossroad between lexicography and terminology work: knowledge organisation and domain labelling') presents the MORDigital project which aims to encode the selected editions of *Diccionario de Lingua Portuguesa*, first published in 1789. The ultimate goal of this work is to promote accessibility to cultural heritage while fostering reusability and to contribute towards a more significant presence of lexicographic digital content in Portuguese through open tools and standards.

Valérie Delavigne ('Des corpus aux bases de données... et retour. Quelle architecture pour une base de données socioterminologiques') tackles the problem of implementing socioterminological data in terminological resources. In particular, the author tries to answer the question of how sociolinguistic diversity can be integrated into a descriptive program, by reviewing the types of variation and exploring an approach to integrating some of these data into a terminographic resource.

In their contribution, Giorgio Maria Di Nunzio, Stefano Marchesin, and Gianmaria Silvello ('A Systematic Review of Automatic Term Extraction: What Happened in 2022?') describe the increasing and variegated interest in the ATE research topic, as one of the key interdisciplinary tasks related to information retrieval, sentiment analysis, and named entity recognition. Starting from a follow-up on the discussions about the pipeline that allows extracting key terms from medical reports, the authors present a systematic review of the recently published papers about ATE.

In their paper entitled 'Terminologie collaborative: analyse d'un projet inter-universitaire outillé en contexte européen', Pascale Elbaz and Elpida Loupaki describe a collaborative terminology project undertaken jointly by three academic institutions (Institute of Intercultural Communication and Management, Aristotle University of Thessaloniki, and University of Padova) as well as the Terminology Coordination Unit of the European Parliament, in the framework of a European cooperation. The authors analyse the theoretical and practical aspects necessary for a successful collaboration as well as an evaluation of the performance of the different tools that were used.

Le travail de Rute Costa, Ana Salgado, Margarida Ramos, Sara Carvalho, Anas Khan, Toma Tasovac, Bruno Almeida, Mohamed Khemakhem, Laurent Romary et Raquel Silva (*A crossroad between lexicography and terminology work: knowledge organisation and domain labelling*) présente le projet MORDigital qui vise à encoder les éditions sélectionnées du *Diccionario de Lingua Portuguesa*, publié pour la première fois en 1789. Le but ultime de ce travail est de promouvoir l'accessibilité au patrimoine culturel tout en favorisant la réutilisation et de contribuer à une présence plus significative du contenu numérique lexicographique en portugais à travers des outils en libre accès et des standards.

Valérie Delavigne (*Des corpus aux bases de données... et retour. Quelle architecture pour une base de données socioterminologiques*) aborde le problème de l'implémentation des données socioterminologiques dans les ressources terminologiques. En particulier, l'auteure vise à répondre à la question concernant l'intégration de la diversité sociolinguistique dans un programme descriptif, en passant en revue les types de variation et en explorant une approche pour intégrer certaines de ces données à une ressource terminographique.

Dans leur contribution, Giorgio Maria Di Nunzio, Stefano Marchesin et Gianmaria Silvello (*A Systematic Review of Automatic Term Extraction: What Happened in 2022?*) décrivent l'intérêt croissant et varié pour le sujet de recherche de l'Extraction Automatique des Termes (EAT), en tant que l'une des tâches interdisciplinaires clés liées à la recherche d'information, à l'analyse des sentiments et à la reconnaissance d'entités nommées. À partir d'un suivi des discussions sur la démarche qui permet d'extraire des termes clés des rapports médicaux, les auteurs présentent une revue systématique des articles récemment publiés sur l'EAT.

Dans leur article intitulé *Terminologie collaborative: analyse d'un projet inter-universitaire outillé en contexte européen*, Pascale Elbaz et Elpida Loupaki décrivent un projet terminologique collaboratif mené conjointement par trois institutions académiques (Institut de management et de communication interculturels, Université Aristote de Thessalonique et Université de Padoue) ainsi que l'Unité de Coordination de la Terminologie du Parlement européen, dans le cadre d'une coopération européenne. Les auteures analysent les aspects théoriques et pratiques nécessaires à une collaboration réussie ainsi qu'une évaluation de la performance des différents outils utilisés.

The contribution ('TBX and lemon: What Perspectives in Terminology?') presented by Silvia Piccini, Federica Vezzani, and Andrea Bellandi proposes a contrastive multilevel analysis between two paradigms for modelling multilingual terminological data: the model proposed in the context of ISO TC 37/SC 3, based on the adoption of the Terminological Markup Framework (TMF)/TermBase eXchange (TBX) standards; and the 'lemon' model and, more generally, the Ontology Web Language adopted in the framework of the Semantic Web. The terminological case study chosen to test the two approaches is represented by the pathology known as 'body dysmorphic disorder', which poses challenges in terms of conceptual and linguistic representation.

The paper by Christophe Roche ('La théorie du concept des normes ISO à l'ère numérique') analyses the contribution of the ISO principles about terminology, as defined by the core standards ISO 1087: 2019 and ISO 704: 2022, to the construction of ontologies for terminological purposes. The author discusses the feasibility of implementing ISO principles in description logic, as the dominant model in knowledge representation.

Finally, the work by Giulia Speranza, Maria Pia Di Buono, and Johanna Monti ('Towards Terminological Resources Tailored to the Users' Needs: Terminology Extraction based on Appositive Constructions') proposes a methodology for identifying and extracting candidate terms from domain corpora. The aim is to collect data suitable for designing different types of terminological resources intended for different target users, that is, translators, interpreters, technical communicators, and laypeople.

## Contribution statement

Federica Vezzani (Writing—original draft, Writing—review & editing) and Giorgio Maria Di Nunzio (Writing—original draft, Writing—review & editing)

## Acknowledgements

We would like to acknowledge the work done by the members of the MDTT 2022 scientific committee as well as all the additional invited reviewers for this Special Issue who have dedicated their time and effort to ensure the quality of the submissions as well as provide feedback and suggestions to authors. We also want to thank the support of the newly born Centre for Computational Terminology (CENTRICO)<sup>5</sup> of the University of Padova whose initiatives are helping to promote research activities on standardized linguistic data modelling and on advanced theoretical models of multilingual specialized terminology and lexicology.

La contribution (*TBX and lemon: What Perspectives in Terminology?*) présentée par Silvia Piccini, Federica Vezzani et Andrea Bellandi propose une analyse contrastive multiniveaux entre deux paradigmes de modélisation de données terminologiques multilingues: le modèle proposé dans le cadre de l'ISO TC 37/SC 3, sur la base de l'adoption des normes *Terminological Markup Framework* (TMF)/*TermBase eXchange* (TBX); et le modèle *lemon* et, plus généralement, l'*Ontology Web Language* (OWL) adopté dans le cadre du Web sémantique. L'étude de cas terminologique choisie pour tester les deux approches est représentée par la pathologie « body dysmorphic disorder », qui pose des défis en termes de représentation conceptuelle et linguistique.

L'article de Christophe Roche (*La théorie du concept des normes ISO à l'ère numérique*) analyse l'apport des principes ISO sur la terminologie, tels que définis par les normes ISO 1087: 2019 et ISO 704: 2022, à la construction d'ontologies à des fins terminologiques. L'auteur discute de la faisabilité de l'implémentation des principes ISO en logique de description, en tant que modèle dominant dans la représentation des connaissances.

Enfin, le travail de Giulia Speranza, Maria Pia Di Buono et Johanna Monti (*Towards Terminological Resources Tailored to the Users' Needs: Terminology Extraction based on Appositive Constructions*) propose une méthodologie pour identifier et extraire des termes candidats à partir de corpus de domaine. L'objectif est de collecter des données adaptées à la conception de différents types de ressources terminologiques destinées à différents utilisateurs cibles, à savoir les traducteurs, les interprètes, les communicateurs techniques et le grand public.

## Author contributions

Federica Vezzani (Writing—original draft, Writing—review & editing) and Giorgio Maria Di Nunzio (Writing—original draft, Writing—review & editing)

## Remerciements

Nous tenons à souligner le travail accompli par les membres du comité scientifique de MDTT 2022 ainsi que tous les relecteurs supplémentaires invités pour ce numéro spécial qui ont consacré leur temps et leurs efforts pour assurer la qualité des soumissions ainsi que pour fournir des commentaires et suggestions aux auteurs. Nous tenons également à remercier le soutien du nouveau centre de terminologie computationnelle (CENTRICO)<sup>5</sup> de l'Université de Padoue, dont les initiatives contribuent à promouvoir les activités de recherche sur la modélisation des données linguistiques normalisées et sur les modèles théoriques avancés de terminologie et de lexicologie spécialisées multilingues.

## Notes

1. <https://eosc.eu>
2. <https://purl.org/fairterm>
3. <http://www.maldura.unipd.it/digital-terminology/en/>
4. <http://mdtr2023.dei.unipd.it/en/>
5. Centro Studi in Terminologia Computazionale—Università degli Studi di Padova. <http://centrico.dei.unipd.it>

Federica Vezzani <sup>1</sup>, Giorgio Maria Di Nunzio <sup>2</sup>

<sup>1</sup>Department of Linguistics and Literary Studies, University of Padova, Via Elisabetta Vendramini 13, Padova 35137, Italy

<sup>2</sup>Department of Information Engineering, University of Padova, Via Gradenigo 6a, Padova 35131, Italy

## References/Références

- Di Nunzio, G. M., Henrot, G. M., Musacchio, M. T., and Vezzani, F. (eds) (2022). 1st International Conference on ‘Multilingual Digital Terminology Today. Design, Representation Formats and Management Systems’, *Proceedings of the 1st International Conference on Multilingual Digital Terminology Today* (Vol. 3161), CEUR. <https://ceur-ws.org/Vol-3161/#preface>
- Vezzani, F. (2021). The FAIRterm resource: Between pedagogical practice and professionalization in specialized translation. *Synergies Italie*, 17: 51–64 (Scopus).
- Vezzani, F. (ed.) (2022). *Terminologie Numérique: Conception, Représentation et Gestion*. Peter Lang International Academic Publishers. <https://doi.org/10.3726/b19407>

## Notes

1. <https://eosc.eu>
2. <https://purl.org/fairterm>
3. <http://www.maldura.unipd.it/digital-terminology/en/>
4. <http://mdtr2023.dei.unipd.it/en/>
5. Centro Studi in Terminologia Computazionale – Università degli Studi di Padova. <http://centrico.dei.unipd.it>

- Vezzani, F. and Di Nunzio, G. M. (2020). Methodology for the standardization of terminological resources: Design of TriMED database to support multi-register medical communication. *Terminology*, 26(2): 265–297 (Scopus). <https://doi.org/10.1075/term.00053.vez>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), Article 1. <https://doi.org/10.1038/sdata.2016.18>

<https://doi.org/10.1093/llc/fqad028>