# Open Gender-sensitive Terminology Resources for the Digital Society: the Digitender Project (Short Paper)

Chelo Vargas-Sierra [1] and M. Isabel Santamaría-Pérez [2]

[1] [2] *IULMA, University of Alicante, Cra. San Vicente del Raspeig s/n, San Vicente del Raspeig, 03690, Spain*

**Abstract**

This paper[1] will present the research project "Open multilingual gender-sensitive terminology resources for the digital society: the DIGITENDER project", led by research groups at IULMA (University of Alicante) and the Computer Linguistics Laboratory at Universidad Autónoma de Madrid. The aim of the DIGITENDER project is twofold. Firstly, it aims to digitize the quality multilingual terminology resources (MTRs) already available at IULMA (specialized bilingual dictionaries of legal, economic, stock market, tourism, etc. terms) compiled some time ago and published by Ariel Editorial, and then publish them on the web in a findable, accessible, interoperable, and reusable format (FAIR). Secondly, women will be visualized in two ways. We will then edit these MTR entries to mark the feminine gender in Spanish in some entries, which we will complete with new subdomains and terms the terminological glossaries already created on women's health issues, which we call WHealth and which include glossaries on menopause, osteoporosis, breast cancer, uterine cancer, etc. DIGITENDER incorporates a gender perspective in the management of terminology on women's health issues and the main pathologies which, because of their prevalence or the fact that they present differently in women to in men, are important to specifically study in women. Our aim is to make visible, through an extensive open terminology repository and terminology datasets in SKOS format, health issues that are particularly related to women. By thus making available to the public some of the issues that arise in this field, such as menopause, menstruation, etc., we also aim to break certain taboos and prototypes that lead to health problems. The project is therefore justified by our will to make available to the public our WHealth glossaries and to develop others from corpora to complete what will be a large repository of gender-sensitive terminology.

**Keywords**

Multilingual terminology, term banks, women's health terminology, terminology datasets

## 1. Introduction

Many studies are currently being carried out on terminology, both its theoretical and methodological aspects. Terminology as a discipline has in fact been receiving increasing attention from theorists and linguists, its importance lying in the fact that it is the field of knowledge where the lexical elements that form part of a structured system of concepts are studied, a system in which particular terms are largely responsible for conveying the knowledge of each scientific, technical, or professional community. Cabré [1] highlights the diverse nature of this discipline, which is revealed in its polyhedral perspective that stems from its foundations, its approaches, and its practical applications, not forgetting the well-known polysemy of the term 'terminology' itself. Rey [2] proposed the name of 'terminography', others refer to this activity as 'terminology work'. Whatever name is given to , it is usually carried out by institutions concerned with defining recommendations for the standardization of terminology, as well as by those institutions, universities, research groups, and organizations that aim to collect together terms for their management and dissemination in term banks, glossaries, etc., with several purposes in mind, such as knowledge transfer, translation, language for specific purposes

2nd International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2023, June 29–30, 2023, Lisboa, Portugal
EMAIL: chelo.vargas@ua.es (A. 1); mi.santamaria@ua.es (A. 2)
ORCID: 0000-0002-4026-4372 (A. 1); 0000-0002-6264-1837 (A. 2)

learning, etc.

A term bank is an essential tool for understanding a term, an acronym, checking a standard denomination, finding an equivalent in another language, and so on. Furthermore, terminology, as a branch of human knowledge, is closely related to other areas of knowledge: applied linguistics, logic, ontology and computer science [1]. It is also linked to all scientific and technical domains; in fact, it is in these domains that its raison d'être originates and is acquires. Indeed, the importance of terminology has grown over time and there is a proliferation of projects focused on improving terminology resources at the national and international level, good examples being IATE, Termium Plus, and TERMCAT.

One of the issues that caught our attention when we considered some of the Spanish terms is that some entries did not include gender markers. What terms convey is knowledge and we may, then, think that they are genderless. This is the case for a large number of high specialized terms that are collected in different resources although many, such as *aborto de repetición* (recurrent miscarriage), *útero dismórfico*, (*dysmorphic uterus*) and *endometriosis*, have a specific gender in Spanish and, as such, obviously cannot be modified. However, there is another type of term, sometimes less specialized, but nonetheless frequent and important in its specific domain, that is only collected and represented in some term banks in its masculine version, when a feminine version does exist, as in the case of Spanish nouns and adjectives that derive from verbs and acquire a meaning of "someone or something that does what is expressed in the root of the verb". These nouns and adjectives with gender motion are usually lemmatized in dictionaries by the masculine, adding the feminine ending (-a, -ra), but this is not always the case, especially in terminological resources. This is the case with 'holder', 'owner', and 'manager', whose Spanish equivalents do not have the feminine marker in some of the terminology banks consulted.

Along with the above-mentioned terminological foundations, our project was shaped by our strong desire to include a gender perspective in our research, but we also wondered how terminologists, translators, and linguists could contribute to the development of more egalitarian societies where the role of women is represented in subject-matter fields equal to that of men, as well as addressing specific women problems. Our research aims to contribute terminologically to a field of knowledge directly related to women and we thus decided that health would be a very interesting field to start with, given the inequalities revealed by analyses carried out by the World Health Organization [3] [4]. The following quote provided the impetus to advance this idea [5: p. 4]:

> The introduction of gender studies into the field of health led to the visibility of many 'new' health problems in women and their study as a relevant and legitimate field of scientific enquiry. Until then, both medicine and public health dealt with women from a narrow perspective focused mainly on reproduction and reproductive pathologies.

According to the above quote, the introduction of a gender perspective in medicine brought to the forefront a number of women's health problems that had not been studied previously, as research had adopted a biased point of view by only paying attention to female-specific diseases, such as reproductive-related problems and pathologies. More recent work on this issue can be found in *Perspectiva de género en medicina* [6] in which we note that this problem is still not completely resolved. Ruíz Cantero [6: p. 10] points out that the concept of "gender bias" in health care appeared for the first time in 1991 in reference to the fact that when faced with the same health needs in men and women, a greater diagnostic or therapeutic effort is made in one sex with respect to the other, which can contribute to inequalities in health. The fact is that the patterns of signs and symptoms, response to treatment and prognosis are mostly formulated based on men; hence, these gender biases can lead to erroneous diagnosis and treatment in women.

In Spain, public health interest in gender inequalities in health began during the early 2000s, promoted, among by, other researchers, Dr. Colomer (2007). The Spanish Society of Public Health promoted the setting up of an observatory (a government agency that monitors women's experiences of health services and promotes their inclusion) discussion forums and the inclusion of gender inequalities in its reports. However, despite these notable and determined efforts, invisibility in this area continues to exist, as is clear from the title of the webinar hosted by the Observatorio de Salud de las Mujeres (Spanish Women's Health Observatory) issued in July 2021 —"Mujeres invisibles para la medicina.

Análisis pre y post pandemia[2]" (Women invisible to medicine. Pre- and post-pandemic analysis) —given by Dr Carme Valls Llobet, Director of the Programa Mujer, Salud y Calidad de Vida (Women, Health and Quality of Life Program) at the Centro de Análisis y Programas Sanitarios (Centre for Analysis and Health Programs).

Taking all the points above into account, our research project incorporates a gender perspective into terminology management (identifying, defining, updating, and maintaining terminology) about women's health issues (WHealth) with the aim of making visible and increasing knowledge about the health problems women usually suffer from, through an extensive open terminology repository. To this end, the terminology record has been defined to contain a gender-specific health care fields in two major groups: signs and symptoms and treatment. In this way, we will make publicly available knowledge about pathologies or disorders that, because of their prevalence in women or because their characteristics are different in women than they are in men, are important to study in women, draw attention to exactly these differences. There are other health issues that are directly related to women, such as menopause and menstruation, or that affect women due to social pressure, such as eating disorders, which we also address in the project, both to disseminate knowledge and increase their visibility, and to break taboos and stigmas. We believe that the creation of multilingual terminology resources for particular fields can help in this task as terminology plays a key role in any field of knowledge and in its dissemination.

DIGITENDER is based on a series of needs that have been observed in the digital transition of multilingual terminological resources created by some members of the IULMA Institute at the University of Alicante and LLI-UAM. More specifically, on reviewing the MTRs available within our research group, we observed that some entries were not gender sensitive. We aim to integrate the gender perspective into certain terminological records; for example, those related to professions.

We have already built glossaries that are available in database format (.mdb), MARTIF (.mtf) and TermBase eXchange (.tbx). Some of these resources are about women's related health issues, such as:

- Assisted reproduction glossary: 437 bilingual terms.
- Menopause glossary: 208 bilingual terms.
- Osteoporosis glossary: 30 bilingual terms.
- Glossary of Breast Cancer: 70 bilingual terms.
- Glossary of Cervical Cancer: 202 bilingual terms.
- Glossary of Eating Disorders: 100 bilingual terms.
- Sexually Transmitted Diseases Glossary: 120 bilingual terms.
- Depression Glossary: 181 + 61 = 242 trilingual terms (plus Catalan).
- Glossary of the Female Reproductive System and its Pathologies: 131.
- Glossary of Rheumatic Diseases: 125 terms.
- Glossary of Pregnancy and Birth Terms: 205 terms.
- In Vitro Fertilisation Glossary: 99 bilingual terms.
- Bilingual Glossary of Healthy lifestyle: 297 bilingual entries.
- Endometriosis Glossary: 102 bilingual terms.
- Glossary of Thyroid Hormonal Diseases: 100 bilingual terms

Our aim is to review and complete these glossaries and develop new corpus-based terminology resources to complete them taking into account the following fields or subdomains:

a. Gynecological health issues and disorders affecting women.
b. Pregnancy-related topics.
c. Infertility-related disorders.
d. Other disorders and diseases that affect only women.
e. Main pathologies which, because of their prevalence or because their characteristics differ from those of men, are important to study in women.

Likewise, DIGITENDER is also based on a series of needs observed in what we hope will be the "digital transition" of the multilingual terminological resources created by some members of IULMA at the University of Alicante. More specifically, when reviewing the MTRs available in our research group, we observed that some terms, such as those mentioned in a previous paragraph, did not include the

[2] Source: https://www.observatoriosaludmujeres.es/mujeres-invisibles-para-la-medicina-analisis-pre-y-post-pandemia/

feminine form. Our aim is to add this gender mark in certain terminological registers; for example, those related to professions (banquero, -a; ejecutivo, -a) and adjectives and nouns with gender motion (propietario, -a; dueño, -a, etc.). The multilingual terminology resources we refer to, all of which have been built by DIGITENDER researchers were published by a renowned publisher, are the following:

a. Dictionary of English-Spanish Legal Terms: 15,000 approx. bilingual entries.
b. Dictionary of Economic, Financial and Commercial Terms: 25,000 bilingual entries.
c. Dictionary of Stock Exchange Terms: 4,000 bilingual entries.
d. Dictionary of Banking Terms, 12,000 bilingual entries.
e. Dictionary of Human Rights Terms: 3.300 bilingual entries.
f. Dictionary of Natural Stone and Allied Industries: 8,000 bilingual entries.

In these resources we estimate that there are approximately 70,000 bilingual English-Spanish entries in total.

## 2. Distribution of terminology resources

Consulting digital linguistic resources is now commonplace and easy. Nevertheless, our specialized dictionaries are out of date, thus we need to digitize them in order to build a term bank and make our resources publicly and freely available in a findable, accessible, interoperable and reusable format [7] [8].

The term "open access" is now applied to refer to data which is configured and distributed in a form that makes it reusable in the context of the semantic web. The Internet, as is well known, has become a powerful medium for communication and research, and an ideal place for the exchange and provision of goods and services of all kinds. Despite its advantages, its enormous potential is partially limited to the human capacity to navigate, without getting lost, through the different sources of information available and the incalculable number of indexed documents.

The exponential increase in the volume of information available on the Internet and the low precision and coverage -or recall- in its retrieval have created technological need in two broad senses: a) to improve the web by extending it with semantic metadata: this extended web is known as the "semantic web" and is constituted as "a universally accessible platform that allows data to be shared and processed by automated tools as well as by people" [9]; and b) to create more agile and efficient information retrieval systems that are capable of understanding and managing information in an "intelligent" way. The semantic web is a set of activities developed within the W3C, the aim of which is to create technologies for the publication of data readable by computer applications. However, for the web of data to become a reality, it is important that the vast amount of data on the web is available in a standard format, accessible and manageable by semantic web tools. Furthermore, this web not only needs access to the data, but also the capacity to facilitate the relationships between them, as the aim is to create an interconnected web of data, as opposed to a simple collection of datasets. This collection of interrelated datasets within the web is also known as "Linked Data" and "Linked Open Data" (LOD)

A great number of resources from different domains have already been published and connected according to certain well-known recommendations, forming what is known as the Linked Open Data cloud (https://lod-cloud.net/). To link terminological resources, we can go to the section of this cloud called "Linguistics" and, within this, to the specific group for "Terminologies, Thesauri and Knowledge Bases", as shown in the following figure:

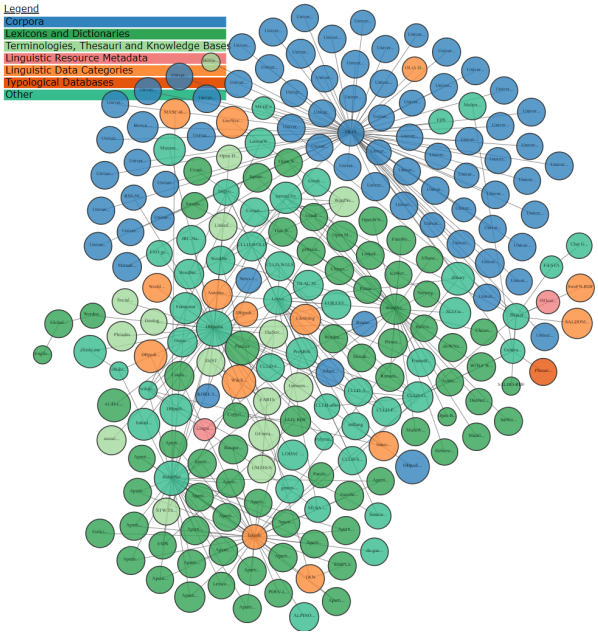


**Figure 1**. Linguistics Linked Open Data Cloud from lod-cloud.net

These initiatives are designed to address the shortcomings of information retrieval on the Internet which require interoperable and multilingual vocabularies.

The Simple Knowledge Organisation System (SKOS) is a data model for sharing and linking knowledge organization systems across the semantic web; it is a "W3C initiative in the form of an RDF application that provides a model for representing the basic structure and content of conceptual schemas" [10], including thesauri and controlled vocabularies. The use of the RDF format for SKOS allows information to be made available in a format readable by software applications, and also to be exchanged and published on the Internet. It is also designed to create new organizational systems or migrate existing ones to the semantic web quickly and easily. An example of this, which has in part inspired our proposal, is the EUROVOC glossary, originally available as an .xls file on the Internet and recently converted to SKOS format. We will analyze our dataset in the different digital formats to provide a tailored and customized solution and enhancement to enable the achievement of the goal related to the conversion of all our terminology datasets to SKOS. These datasets will be distributed through the Zenodo.org platform, within the Universitat d'Alacant Research Data Collection community[3].

## 3. Objectives and methodology

The purpose of DIGITENDER is twofold: on the one hand, it aims to digitize the multilingual terminology resources already available in our research group and to publish them on the Internet in a FAIR format; on the other hand, this project also sets out to make women visible through terminological resources, and as such we will edit specific terminology records in terms of gender marking and will complete the terminology series on women's health issues with new subdomains and terms.

The terminological work that we have been involved in at the IULMA Institute follows the principles postulated by the Communicative Theory of Terminology (CTT) [1], which is pragmatic in nature (principle of adequacy) and corpus-based (principle of natural habitat). Without going into all the specifics of terminology work due to space limitations, the methodology to enhance our multilingual terminology resources will follow the principle of adequacy, which implies that each terminology project must develop both a work strategy and the domain-specific terminology repositories based on variables such as the subject matter, the recipients of the terminology application, the objectives, the context of use of this application and the resources available. In short, a working methodology "configurable by the terminographer" is proposed, i.e., according to the prototypical user of the application we will build and the purposes for which it is intended.

We will carry out a descriptive terminological activity, aimed at compiling and illustrating the

[3] https://zenodo.org/communities/universitat-alacant/?page=1&size=20

specialized items detected in the specialized texts that make up the corpora from which the terms are extracted.

This project will use a cross-disciplinary methodology, drawing on the various disciplines that contribute to its implementation: terminology, computational linguistics, and NLP. The transversal protocol makes use of:

- Methods for descriptive terminology management.
- Methods of corpus design and data mining in "small corpora" and "big data".
- Methods of corpus analysis applied to specialized language.
- Web development technologies and NLP.

All the dataset resulting from this project will be provided in open repositories.

## 4. Concluding remarks

In the previous sections we have outlined the broad lines of the DIGITENDER research project with which we have very recently begun to work. Our project aims to contribute to filling a gap, as we will not only publish resources on women's health issues (those already published as well as new ones to complete the series), but also on the domains of the terminological dictionaries we have already constructed, namely, economics, finance, law, the natural stone industry, human rights, marketing and advertising, banking and the stock exchange. This project is based on the premise that women and men are equal, so this must be reflected in our specialized spoken and written language, in the terminology we use and, therefore, in the linguistic resources we produce, because when something is not named, it does not exist. This applies to some terms that are not listed in their feminine version in Spanish in term banks widely used by translators, linguists and the general public.

In recent years, institutions and companies have adopted semantic web technologies and introduced Artificial Intelligence applications in their processes. We have also observed a strong trend towards open data formats for the publication of datasets, since this translates into a fluid exchange of information with other institutions, organizations or companies. In this context, Linked Data emerges as an interesting initiative that proposes a set of good practices to follow in the effective publishing, sharing and connecting of such data [10]. The use of these technologies is growing all the time as they allow computers and people to work together. Linked data, in this context, is the semantic web's way of linking the various data that are distributed across the web. The WHealth collection will be adapted to be part of the Linked Open Data Cloud, thereby overcoming the weaknesses of the strongly lexically anchored relationships that characterize traditional terminologies.

## Acknowledgements

This publication forms part of the research project "Digitisation, processing and online publication of open, multilingual and gender-sensitive terminology resources in the digital society (DIGITENDER)" (TED2021-130040B-C21), funded by Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación (10.13039/501100011033) and by the European Union "NextGenerationEU"/Plan de Recuperación, Transformación y Resiliencia.

## References

[1] Cabré, M.T. (1999). *La terminología: representación y comunicación*, Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

[2] Rey, A. (1995). *Essays on Terminology*, Amsterdam/Philadelphia: John Benjamins.

[3] World Health Organization. European Observatory on Health Systems and Policies. (2012). *Eurohealth*, 18 (2), World Health Organization. Regional Office for Europe. https://apps.who.int/iris/handle/10665/332876.

[4] World Health Organization. WHO Strategy to integrate gender analysis and actions in the work of WHO. Geneva: World Health Organization. (2007) http://whqlibdoc.who.int/ publications/2009/9789241597708_eng_Text.pdf.

[5] Álvarez-Dardet and Vives-Cases (2012). Three waves of gender and health. *Eurohealth*, 18 (2), World Health Organization. Regional Office for Europe. https://apps.who.int/iris/handle/10665/332876.

[6] Ruiz Cantero, M. T. (coor.) (2019). *Perspectiva de género en medicina*. Barcelona: Fundación Dr. Antoni Esteve. Available at: https://www.esteve.org/libros/perspectiva-de-genero-en-medicina/?doing_wp_cron=1670839661.3348269462585449218750

[7] Erhard Hinrichs and Steven Krauwer. 2014. The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 1525–1531. Reykjavik, Islandia. European Language Resources Association (ELRA).

[8] Meroño-Peñuela, A., de Boer, V., van Erp, M., Zijdeman, R., Mourits, R., Melder, W., Rijpma, A. and R. Schalk. 2020. Ontologies in CLARIAH: Towards interoperability in history, language and media.

[9] Hendler, J., Berners-Lee, T. and Miller, E. (2002): Integrating Applications on the Semantic Web. *Journal of the Institute of Electrical Engineers of Japan*, vol. 122(10), 676-680. http://www.w3.org/2002/07/swint

[10] Pastor Sánchez, J. A., and Martínez Méndez, F. J. (2010). Manual de SKOS (simple knowledge organization system, sistema para la organización del conocimiento simple). *Anales de Documentación*, 13, 285–320. https://revistas.um.es/analesdoc/article/view/107511