Reusability of Biomedical Annotations for Gut-Brain Interplay Information Extraction as Terminological Data

Vanessa Bonato[1,*]

[1] *Department of Linguistic and Literary Studies, University of Padova, Via Elisabetta Vendramini 13 35137 Padova, Italy*

Abstract

In the framework of the CLEF 2025 conference, the GutBrainIE @ CLEF 2025 challenge related to the European-founded project HEREDITARY (HetERogeneous sEmantic Data integration for the guT-bRain interplaY) has been proposed. This Natural Language Processing challenge involves the performance of Named Entity Recognition and Relation Extraction aimed at Information Extraction on a corpus of PubMed abstracts concerning the gut-brain interplay. In this paper, we explore the possibility of reusing entity mentions and relations identified during the gold-standard training dataset annotation process in the form of terminological data in a medical terminology resource.

Keywords

medical terminology, information extraction, biomedical annotation, gut-brain interplay

1. Introduction

Information Extraction (IE) is defined as "the process of automatically extracting structured pieces of information from unstructured or semi-structured text documents" [1]. Within the domain of Information Extraction, two different tasks are typically performed: Named Entity Recognition (NER) and Relation Extraction (RE) [2]. The task of Named Entity Recognition involves "recognizing and categorizing named entities that are presented in a text document" [3]. Instead, Relation Extraction focuses on "identifying the relations between entities from underlying content" [4]. Specifically, the extracted entity relations are of a semantic nature [2].

The evaluation of Information Extraction systems can be systematically performed in campaigns such as CLEF (Conference and Labs of the Evaluation Forum)[2]. In particular, in the context of the CLEF 2025 conference, a task related to the European-founded project HEREDITARY (HetERogeneous sEmantic Data integration for the guT-bRain interplaY)[3] is presented. In fact, the advancement of Information Extraction (IE) systems for the automatic extraction of knowledge on the gut-brain interplay from biomedical texts is one of the wide-ranging objectives of the project. These systems aim to assist healthcare experts in acquiring specialized knowledge extracted from documents concerning the link between gut microbiota and different health conditions, such as mental health-related states [5, 6, 7], Parkinson's disease [8, 9, 10] and other neurological disorders [11, 12, 13] [14].

With the aim of creating a dataset for enhancing targeted IE systems precision, the GutBrainIE @ CLEF 2025 challenge[4] has been proposed. In this Natural Language Processing (NLP) challenge, participants are asked to annotate PubMed abstracts concerning the gut-brain interplay by performing both Named Entity Recognition and Relation Extraction. In this context, the tasks respectively entail: 1) the selection of the text span corresponding to entity mentions in texts and

[1] *4th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2025, June 19-20, 2025, Thessaloniki, Greece.*

[*] Corresponding author.

vanessa.bonato@unipd.it (V. Bonato)

0009-0002-9918-282X (V. Bonato)

[2] https://clef2025.clef-initiative.eu

[3] https://hereditary-project.eu

[4] https://hereditary.dei.unipd.it/challenges/gutbrainie/2025/#

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

their labelling by choosing from a defined list of categories, and 2) the identification of the relations between entity mentions [14]. The two tasks were likewise performed by expert annotators to create the gold-standard training dataset, which will be finally used to train IE systems.

The task of Named Entity Recognition carried out on abstracts by expert annotators differs from the process of term extraction, that is "terminology work that involves the identification and excerption of terminological data by searching through a text corpus" [15]. For instance, in some circumstances, the extracted entity mention cannot be considered a term, defined in terminology science as a "designation that represents a general concept by linguistic means" [15]. For example, the entity mention "oral and gut microbiota" does not represent a term, as two distinct terms designating two different concepts can be identified: 'oral microbiota' and 'gut microbiota'. Nevertheless, in many other cases, the identified entity mention represents a term in the medical terminological domain. This applies to entity mentions such as "major depressive disorder" and "Autism Spectrum Disorder".

In this paper, we assess the extent to which entity mentions and entity relations identified during the annotation process aimed at Information Extraction can be reused in a medical terminology resource, in the form of terminological data concerning the gut-brain axis and gut microbiota-related health conditions.

2. Dataset and Dataset Annotation Description

In this section, we present an overview of the dataset used and provide information about the team of expert annotators. Subsequently, we describe the manual annotation process performed by expert annotators on the set of PubMed abstracts to create the gold-standard training dataset.

2.1. Dataset

The abstracts composing the corpus were extracted from papers systematically selected from the PubMed Electronic Database[5], which is the largest database of biomedical publications. The selection was performed by running two separate queries using the following keywords: 1) "mental health" AND "gut microbiota", and 2) "Parkinson" AND "gut microbiota". Following the exclusion of duplicated documents, the corpus comprehensively amounts to 1663 documents.

The annotation process was carried out by 7 annotators on a total of 403 PubMed abstracts.

2.2. Annotators

The team of annotators is composed of both terminology experts and computer science experts. The group is therefore heterogeneous, specifically comprising three annotators with expertise in terminology and four specialized in computer science.

The terminology work conducted by terminology experts served as the starting point for creating the annotation schema. Indeed, terminology experts are trained to identify terms within textual documents, infer the corresponding general concepts, and detect concept relationships. In particular, the described terminology work was conducted manually, with a view to creating a highly-curated gold-standard annotated dataset. This approach allowed, for instance, to exclusively extract terms pertaining to the medical domain from the abstracts used to create the annotation schema, and to exclude candidate terms from the selection. Terminology experts were then trained by computer science experts to acquire knowledge in both Named Entity Recognition and Relation Extraction.

2.3. Annotation Schema

Following the creation of the corpus, an annotation schema was established by the annotators. The annotation schema defines the set of entity labels that annotators are required to associate to entity

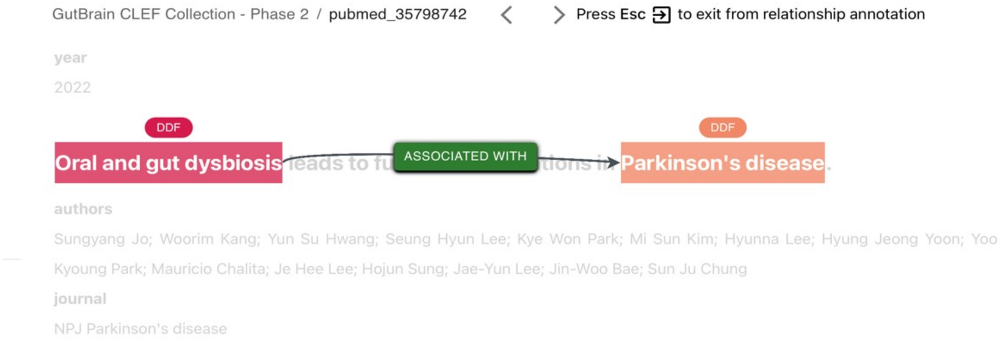
[5] https://pubmed.ncbi.nlm.nih.gov/

mentions identified in the abstracts. The schema also specifies the list of entity relations that link entity mentions, along with the corresponding relation labels.

Concerning entity labels, the schema consists of 14 different categories under which the entity mentions of interest can be classified. Due to space limitations, we will focus on 3 of the 14 labels outlined in the GutBrainIE@CLEF25 Annotation Guidelines[6]. In particular, the labels “Disease, Disorder, or Finding”, “Microbiome” and “Chemical” are particularly relevant to the objectives of the present research.

With reference to entity relations, 22 distinct types can be annotated. For each entity relation, a specific predicate is assigned, considered as a relation label that defines the type of semantic connection between two labeled entity mentions. For example, the entity relation that is established between an entity mention labeled as “Microbiome” and an entity mention labeled as “Disease, Disorder, or Finding” is expressed by using the predicate “is linked to”. In this relation, “Microbiome” is the head entity, while “Disease, Disorder, or Finding” is the tail entity. Another example of entity relation is the relation established between the head entity “Chemical” and the tail entity “Disease, Disorder, or Finding”, whose predicate is “influence”.

2.4. Annotation Process



GutBrain CLEF Collection - Phase 2 / pubmed_35798742

Press Esc to exit from relationship annotation

year

2022

Oral and gut dysbiosis leads to fu… tions in Parkinson's disease.

authors

Sungyang Jo; Woorim Kang; Yun Su Hwang; Seung Hyun Lee; Kye Won Park; Mi Sun Kim; Hyunna Lee; Hyung Jeong Yoon; Yoo Kyoung Park; Mauricio Chalita; Je Hee Lee; Hojun Sung; Jae-Yun Lee; Jin-Woo Bae; Sun Ju Chung

journal

NPJ Parkinson's disease

Figure 1: Screenshot of entity mentions and entity relation annotation.

The annotation process involved the sequential performance of two different tasks for each assigned abstract: Named Entity Recognition (NER) and Relation Extraction (RE). Named Entity Recognition consisted in identifying text spans considered as entity mentions, with the goal of assigning a specific predefined label to each mention. On the other hand, the activity of Relation Extraction concerned the identification of existing relations between pairs of labeled entity mentions explicitly present or inferred within each abstract. The list of defined entity labels and entity relations was provided to annotators through guidelines developed for the challenge.

The annotation workflow consisted of two distinct phases. In particular, in the first phase expert annotators manually annotated a total of 148 abstracts, without pre-annotations for entity mentions and entity labels. The work carried out in this phase led to the identification of 4860 entity mentions and 2360 entity relations. In the second phase, additional 255 abstracts were annotated. In this occasion, however, pre-annotations for entity mentions and entity labels operated by unsupervised algorithms were provided. In the annotated abstracts, 6317 entity mentions and 3045 entity relations were detected.

2.4.1. Mentions

For the selection of text spans constituting entity mentions, specific annotation rules were established. In particular, the following instruction was provided to annotators: “[a]nnotate

[6] For a detailed description of entity labels, relation labels and the annotation rules for entity mentions and entity relations, see https://hereditary.dei.unipd.it/challenges/gutbrainie/2025/#

composite entities as a single entity if they belong to the same category. However, if entities belong to the same category but appear as a sequence, annotate them separately".

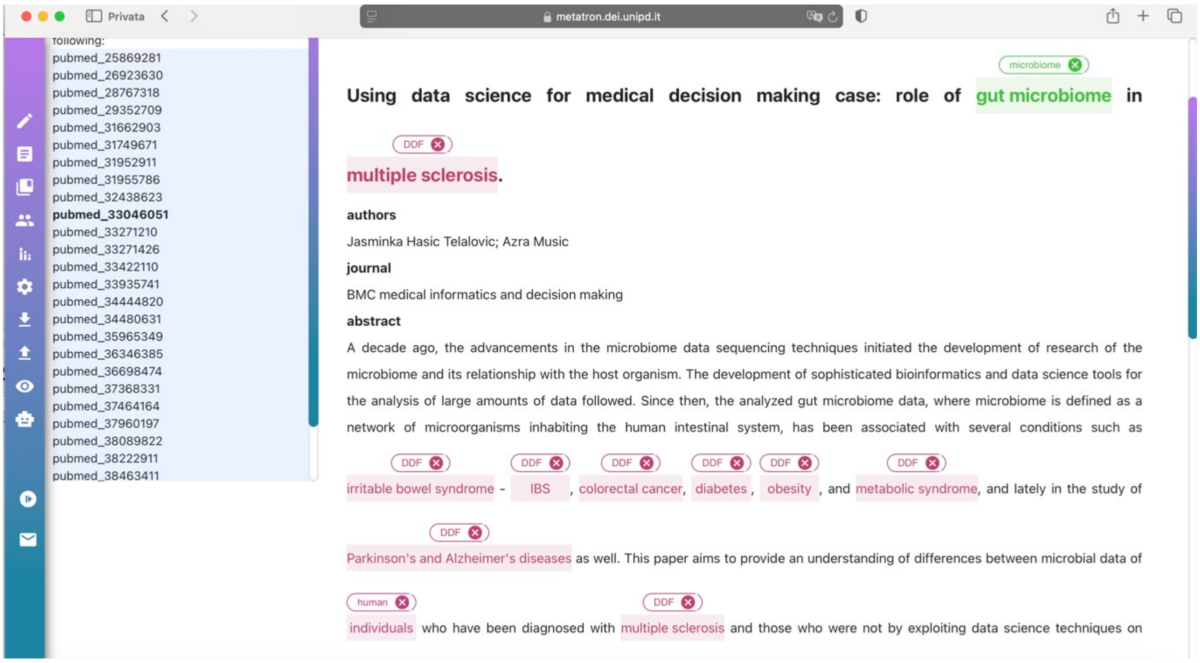
This implies that, for instance, "Parkinson's and Alzheimer's diseases" is considered a single entity mention, due to the fact that the two composite entities belong to the same category. The same reasoning applies to "Oral and gut dysbiosis", "mineralocorticoid and N-methyl-D-aspartate receptors" and to "oral and gut microbiome".

2.4.2. Relations

Concerning the relations established between pairs of labeled entities, a specific instruction was provided in the guidelines. In some cases, indeed, a given predicate may not match the type of semantic connection between entity mentions that can be inferred from the analyzed text. In these circumstances, the predicate "associated with" can be used to signal the existence of a different type of relation between entity mentions. A relation denoted by this predicate has been also used to link entity mentions for which no relation has been established in the guidelines.

For example, provided that an association between two entity mentions labeled "Disease, Disorder, or Finding" explicitly or implicitly emerges from the specific abstract, a relation labeled "associated with" can be annotated. An example of this type of relation can be found in Figure 1, in the context of which the predicate "associated with" is used to specify the link that exists between the entity mentions "Oral and gut dysbiosis" and "Parkinson's disease".

2.5. MetaTron



following:
pubmed_25869281
pubmed_26923630
pubmed_28767318
pubmed_29352709
pubmed_31662903
pubmed_31749671
pubmed_31952911
pubmed_31955786
pubmed_32438623
pubmed_33046051
pubmed_33271210
pubmed_33271426
pubmed_33422110
pubmed_33935741
pubmed_34444820
pubmed_34480631
pubmed_35965349
pubmed_36346385
pubmed_36698474
pubmed_37368331
pubmed_37464164
pubmed_37960197
pubmed_38089822
pubmed_38222911
pubmed_38463411

Using data science for medical decision making case: role of gut microbiome in multiple sclerosis.

authors

Jasminka Hasic Telalovic; Azra Music

journal

BMC medical informatics and decision making

abstract

A decade ago, the advancements in the microbiome data sequencing techniques initiated the development of research of the microbiome and its relationship with the host organism. The development of sophisticated bioinformatics and data science tools for the analysis of large amounts of data followed. Since then, the analyzed gut microbiome data, where microbiome is defined as a network of microorganisms inhabiting the human intestinal system, has been associated with several conditions such as irritable bowel syndrome - IBS , colorectal cancer, diabetes , obesity , and metabolic syndrome, and lately in the study of Parkinson's and Alzheimer's diseases as well. This paper aims to provide an understanding of differences between microbial data of individuals who have been diagnosed with multiple sclerosis and those who were not by exploiting data science techniques on

Figure 2: Screenshot of the MetaTron interface for abstract annotation.

The annotation process was carried out by using the annotation tool MetaTron [16], specifically developed to support biomedical corpora annotation.

The tool enabled annotators to sequentially perform the tasks of Named Entity Recognition and Relation Extraction for each assigned abstract.

3. Can NER and RE Annotations be reused as Terminological Data?

The annotation process aimed at Information Extraction enabled to identify entity mentions and relations in abstracts related to the gut-brain interplay and gut microbiota-related health states. As previously mentioned, however, the task of Named Entity Recognition fundamentally differs from the process of term extraction. Indeed, term extraction is a fundamental step in terminology work, concerned with the extraction of terminological data from document collections [17, 18].

In particular, within the framework of the presented challenge, the need to homogenize the manually performed NER annotations led to the establishment of internal annotation rules to be followed by all annotators. These rules, essential for creating a ground truth for IE systems training, are not necessarily aligned with the terminological approach that is used to extract terms from texts. Considering this, a fundamental distinction characterizes entity mentions and terms. As a matter of fact, in the GutBrainIE dataset, a text span is regarded as a single entity mention when the textual sequence represents composite entities that share the same entity label. Differently, in the terminological domain, the term is the linguistic designation of a concept, that is a "unit of knowledge created by a unique combination of characteristics" [15].

For instance, entity mentions such as "Oral and gut dysbiosis", "mineralocorticoid and N-methyl-D-aspartate receptors" and "oral and gut microbiome", respectively labeled as "Disease, Disorder, or Finding", "Chemical" and "Microbiome", cannot be considered terms. Within the three selected text spans, indeed, six different terms designating six different concepts can be identified: 1) oral dysbiosis, 2) gut dysbiosis, 3) mineralocorticoid receptor, 4) N-methyl-D-aspartate receptor, 5) oral microbiome, and 6) gut microbiome.

As shown in Figure 2, "Parkinson's and Alzheimer's diseases" is also a single entity mention whose composite entities share the entity label "Disease, Disorder, or Finding". In terminology, the text span would not correspond to a term. As a matter of fact, two distinct terms designating two distinct concepts can be identified: 'Parkinson's disease' and 'Alzheimer's disease'.

As can be observed in Figure 2, however, other entity mentions labeled as "Disease, Disorder, or Finding" are identified: "multiple sclerosis", "irritable bowel syndrome", "IBS", "colorectal cancer", "diabetes", "obesity" and "metabolic syndrome". These mentions would be considered medical terms, as each linguistically designates a medical concept. In addition, these terms could be part of lexical networks, where relationships between terms are outlined.

For what concerns entity relations, data emerging from the GutBrainIE gold-standard dataset could also be partially reused in a medical terminology resource in the form of terminological data. Moreover, they can be used to define concept relationships in conceptual systems. For example, the relation established in the guidelines between the head entity "Bacteria" and the tail entity "Microbiome", whose predicate is "part of", matches the part-whole relation, used in terminology as a "concept relation between a comprehensive concept and a partitive concept" [15]. Following this line of reasoning, the concept <microbiome> is the comprehensive concept, that is a "concept in a partitive relation that is viewed as a whole consisting of various parts" [15]. Instead, the concept <bacteria> is the corresponding partitive concept, that represents a "concept in a partitive relation that is viewed as a part of a whole" [15].

On the other hand, it can be observed that generic relations, also defined as "is-a relations", are not considered in the guidelines. In the terminological domain, a generic relation is a "concept relation between a generic concept and a specific concept where the intension of the specific concept includes the intension of the generic concept plus at least one additional delimiting characteristic" [15].

Another observation concerns the predicate "associated with", used to link entity mentions when predefined relation labels do not accurately express the relation established in a specific abstract. This predicate exclusively suggests that a link exists between two entity mentions, without specifying the particular kind of entity relation that is established. By way of exemplification, in Figure 1, the predicate "associated with" marks the relation between "Oral and gut dysbiosis" and "Parkinson's disease". In terminology work, associative relationships are used to link concepts that are not involved in generic relations or part-whole relations. However, in conceptual systems, it would be necessary to precisely indicate the kind of associative relationship established between concepts. In this sense, an additional fine-grained level of analysis from a semantic viewpoint should be considered in entity relation labeling, with a view to precisely systematizing conceptual knowledge.

4. Conclusions

In this paper, we investigated the possibility of reusing data stemming from the manual annotation of the GutBrainIE gold-standard training dataset for Information Extraction in the form of terminological data.

Our analysis highlighted that entity mentions and relations can be partially reused as terminological data related to the gut-brain interplay in a medical terminology resource, as well as in domain-specific lexical networks and conceptual systems. In particular, a selection should be performed to identify entity mentions that are considered terms in the medical terminological domain. Moreover, for terminological conceptual analysis, it would be essential to integrate information on generic relations and to further specify the predicates that denote associative relations between entity mentions.

As future work, we aim to compare the gold-standard annotated dataset with the output generated by automatic term extractors, in terms of both precision and recall. Furthermore, we aim to provide further information about the terminology work that served as the foundation for the creation of the annotation schema. Finally, we will analyze additional entity mentions and entity relations included in the annotated dataset to further investigate how NER and RE annotations can be reused as terminological data in a medical terminology resource.

Acknowledgments

This work is partially supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074.

Declaration on Generative AI

During the preparation of this work, the author(s) used **ChatGPT** in order to: **Grammar and spelling check**, **Paraphrase and reword**. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. Chiticariu, M. Danilevsky, H. Ho, R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, S. Vaithyanathan, H. Zhu, Web Information Extraction, in: L. Liu, M.T. Özsu (Eds.), Encyclopedia of Database Systems, Springer, New York, NY, 2018, pp. 4620-4629. doi:10.1007/978-1-4614-8265-9_459.

[2] Z. Nasar, S. W. Jaffry, M. K. Malik, Named Entity Recognition and Relation Extraction: State-of-the-Art, ACM Comput. Surv. 54, 1, (2021) 20. doi:10.1145/3445965.

[3] B. Jehangir, S. Radhakrishnan, R. Agarwal, A survey on Named Entity Recognition − datasets, tools, and methodologies, Natural Language Processing Journal 3 (2023) 100017. doi:10.1016/j.nlp.2023.100017.

[4] X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, R. Xu, A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers, ACM Comput. Surv., 56(11) (2024) 293. doi:10.1145/3674501.

[5] E. Gulas, G. Wysiadecki, D. Strzelecki, O. Gawlik-Kotelnicka, M. Polguj, Can microbiology affect psychiatry? A link between gut microbiota and psychiatric disorders, Psychiatria Polska, 52(6) (2018) 1023-1039. doi:10.12740/PP/OnlineFirst/81103.

[6] A. Dziedzic, K. Maciak, K. Bliźniewska-Kowalska, M. Gałecka, W. Kobierecka, J. Saluk, The Power of Psychobiotics in Depression: A Modern Approach through the Microbiota-Gut-Brain Axis: A Literature Review, Nutrients, 16(7) (2024) 1054. doi:10.3390/nu16071054.

[7] S.-Y. Kim, S.-Y. Woo, S. Raza, D. Ho, S. W. Jeon, Y. Chang, S. Ryu, H.-L. Kim, H.-N. Kim, Association between gut microbiota and anxiety symptoms: A large population-based study examining sex differences, J Affect Disord., 333 (2023) 21-29. doi:10.1016/j.jad.2023.04.003.

[8] Y. Liang, L. Cui, J. Gao, M. Zhu, Y. Zhang, H.-L. Zhang, Gut Microbial Metabolites in Parkinson's Disease: Implications of Mitochondrial Dysfunction in the Pathogenesis and Treatment, Mol Neurobiol., 58(8) (2021): 3745-3758. doi:10.1007/s12035-021-02375-0.

[9] F. Bai, L. You, H. Lei, X. Li, Association between increased and decreased gut microbiota abundance and Parkinson's disease: A systematic review and subgroup meta-analysis, Experimental Gerontology, 191:112444 (2024). doi:10.1016/j.exger.2024.112444.

[10] M. Feng, Z. Zou, P. Shou, W. Peng, M. Liu, X. Li, Gut microbiota and Parkinson's disease: potential links and the role of fecal microbiota transplantation, Frontiers in Aging Neuroscience, 16:1479343 (2024). doi:10.3389/fnagi.2024.1479343.

[11] P. Oroojzadeh, S. Y. Bostanabad, H. Lotfi, Psychobiotics: the Influence of Gut Microbiota on the Gut-Brain Axis in Neurological Disorders, J Mol Neurosci 72(9) (2022): 1952-1964. doi:10.1007/s12031-022-02053-3.

[12] M. You, N. Chen, Y. Yang, L. Cheng, H. He, Y. Cai, Y. Liu, H. Liu, G. Hong, The gut microbiota-brain axis in neurological disorders, MedComm, 5(8):e656 (2024). doi:10.1002/mco2.656.

[13] Y. He, K. Wang, N. Su, C. Yuan, N. Zhang, X. Hu, Y. Fu, F. Zhao, Microbiota-gut-brain axis in health and neurological disease: Interactions between gut microbiota and the nervous system, Journal of Cellular and Molecular Medicine, 28(18):e70099 (2024). doi:10.1111/jcmm.70099.

[14] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodriguez Ortega, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, L. Menotti, G. Silvello, G. Paliouras, BioASQ at CLEF2025: The thirteenth edition of the large-scale biomedical semantic indexing and question answering challenge, in: Proceedings of the 47th European Conference on Information Retrieval (ECIR 2025).

[15] ISO 1087:2019, Terminology work and terminology science – Vocabulary, 2019. URL: https://www.iso.org/standard/62330.html.

[16] O. Irrera, S. Marchesin, G. Silvello, MetaTron: advancing biomedical annotation empowering relation annotation and collaboration, BMC Bioinformatics 25, 112 (2024). doi:10.1186/s12859-024-05730-9.

[17] F. Vezzani, G. M. Di Nunzio, S. Silecchia, La fraseologia dei trattati internazionali di disarmo: la risorsa terminologica DITTO, Umanistica Digitale 14 (2022) 91-117. doi:10.6092/issn.2532-8816/14796.

[18] A. Giovagnoli, F. Vezzani, G. M. Di Nunzio, La comunicazione medico-paziente: analisi terminologica e di semplificazione nel dominio dell'oncologia femminile, Umanistica Digitale 17 (2024) 143-164. doi:10.6092/issn.2532-8816/19265.