

HEREDITARY

HetERogeneous sEmantic Data integration for the guT-bRain interplaY

Deliverable 3.4

MEDICAL TERMINOLOGY

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



**Funded by
the European Union**

EXECUTIVE SUMMARY

This report outlines a methodology that addresses both conceptual and linguistic dimensions of terminology with the prime aim to enhance knowledge representation and promote informed communication within the medical domain, especially for the diseases under study in HEREDITARY.

Foundational to this initiative is an approach that merges terminological theory with practical application. The theoretical underpinnings emphasize the dual nature of terminology: it operates both as a conceptual structure that reflects domain knowledge and a linguistic system of specialized terms. This conceptual-linguistic synergy ensures terminological accuracy, consistency, and clarity, ultimately improving the quality of healthcare information transfer.

To demonstrate this dual-dimension approach we will go through an in-depth exploration of the gut-brain axis in existing biomedical terminological resources. The methods aim to illustrate how conceptual and linguistic structuring supports better domain understanding, corpus building, and expert engagement. Domain-corpus building and subsequent exploitation is a gateway to domain knowledge verbally expressed in texts written by experts. Hence, documenting the criteria, typologies, and metadata of gathered texts ensures a solid empirical foundation for terminology extraction. Various methods are presented for automatic and semi-automated term extraction. Tailored for the medical sector, these approaches address complexity and domain specificity, thus improving the precision and relevance of the extracted terms.

Validating the terminology to ensure both linguistic accuracy and conceptual integrity is a core activity. By clarifying roles, processes, and the importance of citizen engagement, this validation step ensures that the resulting terminology is both authoritative and accessible to various user communities.



DOCUMENT INFORMATION

Deliverable ID	D3.4
Deliverable Title	Medical terminology
Work Package	WP3
Lead Partner	UNL
Due date	31.12.2024
Date of submission	18.12.2024
Type of deliverable	R
Dissemination level	PU

AUTHORS

Name	Organization
Rute Costa	UNL
Federica Vezzani	UNIPD
Giorgio Maria Di Nunzio (Reviewer)	UNIPD
Margarida Ramos	UNL
Raquel Silva	UNL
Sara Carvalho	UNL
Vanessa Bonato	UNIPD
Matilde Canelas	UNL
Svetla Boytcheva	ONTO
Anna Romanovych (Contributor)	UNIPD

REVISION HISTORY

Version	Date	Author	Document history/approvals
V0.1	31.05.24	Rute Costa	First draft proposal / outline
V0.2	06.06.24	Giorgio Maria Di Nunzio	Adjustments to the outline.
V0.3	29.07.24	Margarida Ramos	Update: (1) Document information and (2) Authors.
V0.4	10.10.24	Margarida Ramos	Sections headings development: Sub-sections 3, 5, 6, 7 and 8.

Version	Date	Author	Document history/approvals
V0.5	25.10.24	Giorgio Maria Di Nunzio, Federica Vezzani	Sections headings updates (Sec 6, 7, 9).
V0.6	18.11.2024	Federica Vezzani	Draft content for Section 9 + bibliography for Section 9
V0.7	18.11.2024	Vanessa Bonato, Federica Vezzani	Draft content for Section 7 + bibliography for Section 7
V0.8	19.11.2024	Giorgio Maria Di Nunzio	Added content for Section 6.
V0.9	22.11.2024	Giorgio Maria Di Nunzio	Added references and fixed typos in Section 6.
V0.10	24.11.2024	Sara Carvalho	Draft content for Section 3 + bibliography for Section 3 + 4 annexes
V0.11	25.11.2024	Raquel Silva	Draft content for Section 8 + bibliography for Section 8
V0.12	25.11.2024	Margarida Ramos, Matilde Canelas	Added content for: Section 5 + bibliography. Section 6, Section 6.2.
V0.13	26.11.2024	Rute Costa	Draft content for Section 2 + bibliography for Section 2
V0.14	26.11.2024	Margarida Ramos, Matilde Canelas	Added content for: Section 6, Section 6.5 + bibliography
V0.15	27.11.2024	Margarida Ramos	Proposal of a separate Section for the content 'Ongoing work in HEREDITARY'
V. 16	27.11.2024	Svetla Boytcheva	Added content for Section 3.
V0.17	28.11.2024	Margarida Ramos	<ol style="list-style-type: none"> 1. Update of content in Section 4 'CORPUS' and in Section 5.5– Results and data analyses removed. 2. Main document: <ol style="list-style-type: none"> a. Text formatting. b. Section 4 – Heading 'Domain description' deleted. c. Update of Section numbering. d. Update of table of contents.
V0.18	28.11.2024	Rute Costa	Draft content for: <ul style="list-style-type: none"> • Section 1 – Introduction. • Section 9 – Conclusion. Added content to Section 6.3.
V0.19	28.11.2024	Rute Costa, Margarida Ramos	Revision and formatting of the main document.
V0.20	29.11.2024	Margarida Ramos	Correction of text formatting and number of annexes.

Version	Date	Author	Document history/approvals
V0.21	04.12.2024	Giorgio Maria Di Nunzio	Review of the document
V0.22	06.12.2024	Rute Costa, Margarida Ramos	Review of V0.21
V0.23	10.12.2024	Giorgio Maria Di Nunzio	Final check
V0.24	10.12.2024	Rute Costa, Margarida Ramos	Final check (2)
V0.25	12.12.2024	Anna Romanovych	Check on the overall formatting of the document.
V0.26	16.12.2024	Margarida Ramos; Rute Costa	Executive summary update and references systematization according to the template.
V1.0	18.12.2024	Anna Romanovych	Preparation of the final version of the deliverable

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Contents

1	Introduction	11
2	Terminology: linguistic and conceptual dimensions	13
2.1	Theoretical framework	13
2.2	Definitions of key concepts	14
3	The Gut-brain interplay and its representation in current biomedical terminological resources	16
3.1	Terminology and healthcare in the 21st century	16
3.2	Representation of the gut-brain interplay in current biomedical terminological resources	17
3.2.1	<Gut-Brain Axis>	18
3.2.2	<Neurodegenerative Diseases>	19
3.3	Discussion and next steps	20
4	Domain-specific corpus: tools and methods	22
4.1	Corpus compilation process: criteria, typology, and goals	22
4.1.1	Full papers: a paramount criterion for terminological data extraction	22
4.1.2	Subcorpus: a collection of abstracts	22
4.1.3	Challenges and opportunities	23
4.2	Documentation: text types, criteria, and metadata	23
4.2.1	Text-related and authorship-ID information	23
4.2.2	Statistical data of the domain-specific corpus for HEREDITARY	23
4.3	Corpus management and future directions	24
4.3.1	Project-related data sources: legal and ethical aspects	24
5	Terminology extraction	25
5.1	Methodologies and approaches	26
5.2	Tools and software for term extraction	26
5.3	Evaluation metrics and benchmarking	28
5.4	Specialized term extraction: medical domain	29
5.5	Semi-automated terminology extraction: tools and methods	31
5.5.1	Terminological data systematization	32
5.5.2	Data analysis	33
5.5.3	Lexical markers pointing at lexical-semantic relations	33
5.5.4	Rule-based methods for lexical-semantic relations identification	34

6	Conceptual and lexical relation typologies	35
6.1	Introduction to conceptual and lexical relationship typology	35
6.2	Typology of concept relationships	35
6.2.1	ISO 1087's approach to concept relationship typology	35
6.2.2	Nuopponen's approach to conceptual relationship typology	36
6.3	Typology of lexical relationships	37
6.3.1	Synonymy and near-synonymy	37
6.3.2	Hypernym and hyponym	37
6.3.3	Meronym and holonym	37
6.4	Conceptual and lexical relationship typology in the medical domain	37
7	Health terminology validation	39
7.1	Key aspects promoting validation	39
7.1.1	Advances in health	39
7.1.2	Improving health communication	40
7.1.3	Increasing Health Literacy (HL)	40
7.2	Terminology validation processes	42
7.2.1	Validation and verification	42
7.2.2	Mediation process	42
7.2.3	Framework for validation	43
7.3	Linguistic guidelines for validation	44
7.3.1	Objectives for validation	44
7.3.2	Stakeholders and end users	44
7.3.3	Selection of terms and concepts for validation	45
7.3.4	Orientations for terminological validation	45
8	Design of a FAIR terminology resource	47
8.1	Introduction to FAIR principles	47
8.2	Overview of the FAIR terminology paradigm	47
8.3	Current state of research	49
8.3.1	Data entry interface	50
8.3.2	Data consultation interface	51
9	Conclusions	53
10	Annexes	54
	REFERENCES	55

List of Tables

Table 1. Requirements for terminology validation.....	43
---	----

List of Figures

Figure 1. ISO TC 37 1087:2019.....	14
Figure 2. FAIRterm 2.0, concept level (screenshot).	50
Figure 3. FAIRterm 2.0, language level (screenshot).	51
Figure 4. FAIRterm 2.0, data consultation interface (screenshot).	52

1 Introduction

Medical terminology used by healthcare professionals integrates conceptual and linguistic systems to structure and organize knowledge, facilitating communication among experts, patients, and non-experts. These systems provide a robust framework for effective healthcare communication, ensuring precise terminology that ideally fosters a shared understanding.

Within the framework of the Hereditary project, our methodology is grounded in the dual dimension of terminology (cf. Section 2) to ensure the quality of terminological data at both the conceptual and linguistic levels, which form the foundation for future terminological work in this project.

Aside from the introduction and conclusion, the report comprises seven distinct sections. Section 2 – Terminology: linguistic and conceptual dimensions – focuses on the theoretical framework that underpins the report's content. It explores the dual dimension of terminology and provides definitions of the essential concepts used in this field of work.

Section 3 deals with The Gut-brain interplay and its representation in current biomedical terminological resources and focuses on describing an essential step in terminology work: the familiarization with the domain under study. As referred to in section 2, acknowledging that Terminology has a double dimension has theoretical and methodological implications. For doing so, one went through the <Gut-brain> representation of the gut-brain interplay in current biomedical terminological resources such as BioPortal, HeTOP, Athena, and the Ontology Lookup Service (OLS). These resources collectively facilitate extensive biomedical data exploration and interoperability.

The analysis of this tool supports terminologists in their work, as systematization and organization are fundamental to effective terminology management. These processes provide valuable assistance by: i) improving comprehension of the domain under study; ii) streamlining corpus-related tasks; iii) enabling productive interactions with subject matter experts and stakeholders; and iv) supporting validation efforts, particularly in the development of precise natural language definitions.

In section 4, dedicated to the corpus, we aim to illustrate the methods employed for compiling a corpus, including the criteria, typology, and objectives. Additionally, we document the corpus development process by organizing the collected texts into categories and incorporating metadata.

Section 5 focuses on terminology extraction, presenting various methodologies and approaches. These include linguistic methods, statistical techniques, machine learning-based strategies, and deep learning approaches. Additionally, we discuss several tools and software used for term extraction, providing detailed descriptions. The section also addresses evaluation metrics and benchmarking techniques for assessing the performance of automatic term extraction (ATE) systems.

Within the medical domain, specialized methods, tools, and benchmarks are employed to tackle the unique challenges arising from the complexity and specificity of medical language. By focusing on these domain-specific resources, the study seeks to improve the efficiency and accuracy of term extraction processes in medical contexts. We further advance our work by demonstrating the application of semi-ATE to medical corpora for identifying and systematizing lexical semantic relations.

In Section 6, we shift focus to conceptual and lexical relation typologies, with particular emphasis on Nupponen's approach to conceptual relations in the medical domain. We move into medical resources such as SNOMED CT, ICD, and UML.

In Section 7, "Health Terminology Validation," we outline the methodology for validating terminology across both linguistic and conceptual dimensions to enhance health communication and improve health literacy. We differentiate between validation and verification processes, discuss the mediation process, and present a framework for terminology validation. In this context, we present a typology of domain experts and conclude by highlighting the importance of citizen engagement.

In Section 8, we provide a detailed overview of the FAIR Terminology resource, emphasizing the FAIR principles as essential for its design. We discuss the current state of research and then present the FAIR Term Web application in detail, including both the data entry interface and the data consultation interface.

The report concludes with a comprehensive bibliography and several appendices, following the concluding remarks.

2 Terminology: linguistic and conceptual dimensions

2.1 Theoretical framework

According to ISO TC 37 ISO 1087-1:2019, terminology science explores “terminologies, aspects of terminology work, the resulting terminology resources, and terminological data” (p. 2). Terminology science, or simply terminology, is the study of concepts and the terms used to represent them. Concepts are part of a conceptual system, while terms belong to a lexical system. One of the fundamental tasks of a terminologist, or more broadly, of any domain expert, is to relate terms to concepts and concepts to terms. The goal of making concepts explicit and understandable, both among experts and non-experts, is essential in any context of knowledge organization and communication. Defining concepts is a core activity in terminology work, with the aim of conveying knowledge in the clearest and most unambiguous way possible. A definition is a statement or formula that establishes a stable relationship between a concept and its associated terms, based on a consensus among those involved in the communication and/or organization of knowledge. It offers a clear and precise understanding that differentiates the concept from other similar concepts, and the terms from other related terms, helping to ensure clarity and prevent ambiguity. Definitions can vary in complexity, from simple, straightforward explanations to more detailed and technical descriptions, depending on the context in which they are used.

Concepts and terms belong to two distinct levels of analysis, corresponding to the two dimensions of terminology: the conceptual dimension and the linguistic dimension. In the conceptual dimension, the focus is primarily on concepts, conceptual relations, and formal definitions. In the linguistic dimension, attention shifts to terms, lexical relations, and definitions expressed in natural language. This suggests that the methodological approaches to terminological data may be threefold: (i) semasiological, (ii) onomasiological, and (iii) mixed. In terminology work, the semasiological approach, as outlined by Zauner (1902), begins with the linguistic unit as the starting point. The main research questions are: (i) which concept (Begriff) is associated with this unit, and (ii) what meaning (Bedeutung) is conveyed by it? In contrast, the onomasiological approach starts with the concept, and the key research question is: which designation is available for this concept? The mixed approach, which we advocate, involves the integration of both methods, as the workflow necessitates transitioning between terms and concepts at various stages.

Working in multilingual and multicultural contexts presents even greater challenges. Multilinguality, and by extension cross-culturality, represents an important dimension in terminology science. It entails understanding, developing, and managing terms and concepts across different languages and cultural contexts within specific subject fields. This perspective recognizes the intrinsic connection between language and culture, emphasizing that terms in one language often cannot be directly translated or aligned with those in another without accounting for cultural nuances. By incorporating multilingual and cross-cultural dimensions, terminology science enhances the understanding of how language and culture interact to shape human knowledge.

This report introduces the methodological approaches that will underpin the work conducted in Task 3.1.

2.2 Definitions of key concepts

To clarify the key meta terminology used in this report, we provide the definitions of "concept," "term," and "definition" as outlined in ISO 1087-1: 2019. These definitions are as follows:

3.2.7 concept

unit of knowledge created by a unique combination of *characteristics* (3.2.1)

Note 1 to entry: Concepts are not necessarily bound to particular *natural languages* (3.1.7). They are, however, influenced by the social or cultural background which often leads to different categorizations.

Note 2 to entry: This is the concept 'concept' as used and designated by the term "concept" in *terminology work* (3.5.1). It is a very different concept from that designated by other domains such as industrial automation or marketing.

3.4.1 designation

designator

representation of a *concept* (3.2.7) by a sign which denotes it in a *domain* (3.1.4) or *subject* (3.1.5)

Note 1 to entry: A designation can be linguistic or non-linguistic. It can consist of various types of characters, but also punctuation marks such as hyphens and parentheses, governed by domain-, subject-, or language-specific conventions.

Note 2 to entry: A designation may be a *term* (3.4.2) including *appellations* (3.4.3), a *proper name* (3.4.4), or a *symbol* (3.4.5).

3.3.1 definition

representation of a *concept* (3.2.7) by an expression that describes it and differentiates it from related concepts

Figure 1. ISO TC 37 1087:2019

Nuopponen (1994) developed the concept of "concept relations" in her doctoral thesis, *Concept Systems for Terminological Analysis*, where she introduced a classification system to improve terminological analysis. In subsequent works (2018, 2022), she defined "conceptual relations" as the connections between concepts that support the organization, analysis, and definition of domain-specific knowledge. These relations are crucial for building terminologies and structuring specialized knowledge by logically and hierarchically linking concepts. Concept relations help define the interdependencies or interactions between concepts, which can include different types of relationships (e.g. hierarchical relations, associative relations, causal relations).

Regarding lexical relations, several authors with differing theoretical perspectives have contributed to the understanding of how lexical units are connected. For instance, Cruse (2000), Lyons (1977), and Fellbaum (1998) all explore the ways in which lexical relations connect words through their meanings, forms, or syntactic roles. These relations can

take various forms, such as hierarchical (e.g., hyponyms and hypernyms) or part-whole associations (e.g., meronyms and holonyms).

The theoretical distinction between conceptual and lexical relations underscores the dual nature of terminology. When approaching terminology from a semasiological perspective, lexical relations within texts often reveal underlying conceptual relations. While conceptual relations themselves are not directly expressed in texts, they are implicitly conveyed through the lexical relations that connect words or terms. This distinction highlights the importance of understanding both the semantic connections between lexical units or terms (lexical relations) and the broader conceptual structures they represent (conceptual relations) (cf. Nuopponen, 2014)

As outlined in ISO/FDIS 5078:2024, "terminology extraction begins with the collection of a text corpus, based on the project's objectives." This process is further described as the "identification and extraction of candidate terminological data" (ISO/FDIS 5078:2024). The methodology presented in this report is informed by analyses of biomedical resources and employs a corpus-driven approach, utilizing a mixed-methods strategy. This combined approach ensures a comprehensive extraction process, drawing on both qualitative and quantitative data to identify relevant terminological elements effectively.

3 The Gut-brain interplay and its representation in current biomedical terminological resources

3.1 Terminology and healthcare in the 21st century

Characterized by significant breakthroughs and driven by unprecedented technological innovation, today's healthcare landscape is vibrant and multifaceted, striving to cater to the needs of various (and increasingly participative) stakeholders, namely in what concerns equitable access to accurate and clear information. Adaptability to this rapid pace is paramount, ensuring that medical terminology remains current and reflects such advancements.

Current biomedical terminological systems¹ (e.g. classifications, thesauri, terminologies, ontologies, etc.) seek to address such diverse needs while aiming towards a semantically interoperable² ecosystem. Built upon a set of principles put forward towards the end of the 1990s (cf. Cimino, 1998; Chute, 1998; Rector, 1999), today's biomedical terminological systems are mostly concept-oriented, anchored in (what would be ideally) systematic - and standardized - representations of conceptual and linguistic information.

While the ultimate goal of full semantic interoperability is still to be attained, important steps have been taken towards more consistent mapping - and linking - of biomedical datasets containing both conceptual and linguistic information. However, such resources must account for the dynamic nature of medical language, where there is often no straightforward one-to-one relation between concept and term (and vice versa). Moreover, clinical concepts are often multidimensional, as well as deeply rooted in postulates that vary across cultures.

In order to capture this complexity, more flexible approaches to concept representation, organization, and sharing have recently been introduced into the structure of such biomedical terminological resources, namely by enabling polyhierarchy and relying on compositional rules (cf. Carvalho, 2018). The resources in question are also increasingly incorporating linguistic information (such as textual definitions) and becoming more and more multilingual.

While terminology work can greatly benefit from such resources as a way to gather domain knowledge and help support corpus-related tasks, it can also contribute towards their development and/or enrichment, supported by a theoretical and methodological framework such as the one put forward in Section 2. The relevant (and more recent) interplay involving Terminology and ontologies - thereby enabling a terminological concept system to be represented as a formal ontology - has been an example of this reciprocal and beneficial influence, contributing to more consistent medical knowledge representation (cf. Carvalho, 2018).

¹ We understand terminological systems in this subject field as resources whose aim is “to organize the relationships between terms and concepts in the biomedical domain with, when appropriate, any associated rules, relationships, definitions, and codes” (Duclos et al., 2014: 22).

² We refer here to the definition of semantic interoperability as the “ability for data shared by systems to be understood at the level of fully defined domain concepts” (ISO/TR 14639-1:2012).

To illustrate this approach in the scope of the HEREDITARY project (and more specifically of T3.3), the following subsections aim to describe how the gut-brain interplay is currently represented in a set of concept-oriented biomedical terminological resources. More specifically, it aims to collect and organize the available information, both from a conceptual and linguistic standpoint, regarding some of the key concepts of the HEREDITARY project. While the primary focus is on <Gut-Brain Axis>³, the project's central concept, one of HEREDITARY's disease groups - **neurodegenerative diseases** - has also been selected as a case study⁴.

3.2 Representation of the gut-brain interplay in current biomedical terminological resources

The data was collected via BioPortal⁵ and HeTOP⁶. While the former, developed by the Stanford Center for Biomedical Informatics Research, is widely regarded as one of the world's most comprehensive repositories of biomedical ontologies (containing more than 1100 ontologies), the latter, developed by the CISMef team, from the Rouen University Hospital, provides access to about 100 terminologies/ontologies in the healthcare domain, in over 50 languages. Another relevant biomedical terminology resource is Athena – OHDSI Vocabularies Repository⁷, that provides access to 156 medical ontologies and vocabularies, and mappings to their standard Observational Medical Outcomes Partnership (OMOP) equivalent. Athena is developed and maintained by the Observational Health Data Sciences and Informatics (OHDSI). Athena is a valuable terminological resource, because it provides three categories of concepts: classification, standardized and non-standardized. Ontology Lookup Service (OLS)⁸ is maintained by the Samples, Phenotypes and Ontologies Team (SPOT) at EMBL-EBI. OLS contains 266 ontologies.

Both tools were used in a complementary way, allowing the creation of a database with the following data categories: a) resource name; b) concept ID; c) preferred name/label⁹; d) synonym(s)¹⁰ or also called alternative labels; e) textual definitions; f) immediate superordinate concept (when applicable); g) top-level concept (when applicable). In what concerns the conceptual dimension, an attempt to organize and represent the collected data was undertaken, despite the fact that, as mentioned before, the terminological resources in question have distinct purposes and follow, therefore, different philosophies and principles. Nevertheless, it is believed that such a systematization can play an important role in terminology work, particularly in helping the terminologist to i) better grasp the domain under analysis; ii) set up and optimize corpus-related tasks (cf. Section

³ To ensure a systematic representation, concepts are depicted between chevrons and with an initial capital letter. Terms are usually represented in lower case and between inverted commas. Since the collected linguistic expressions in the consulted biomedical terminological resources may not be terms *per se*, they will be represented in lower case and italics.

⁴ Future work involves replicating this analysis for the remaining HEREDITARY disease groups.

⁵ <https://bioportal.bioontology.org/> and Grosjean et al. (2011).

⁶ <https://www.hetop.eu/hetop/> and Whetzel et al. (2011).

⁷ <https://athena.ohdsi.org/search-terms/terms/37110787>

⁸ <https://www.ebi.ac.uk/ols4>

⁹ This follows the nomenclature used by Bioportal and HeTOP, respectively.

¹⁰ *Ibid.*

4); iii) prepare any foreseen interactions with subject field experts and/or other relevant stakeholders; iv) support validation processes (cf. Section 7), particularly as far as natural language definitions are concerned. SNOMED CT¹¹ also provides rules for post-coordination, which allows the creation of new, more specific concepts that are not explicitly present in the ontology¹².

For the purpose of this deliverable, the information pertaining to the linguistic dimension was gathered in English only. Multilingual data collection and systematization have been planned at later stages of the work developed in T3.3.

3.2.1 <Gut-Brain Axis>

While the concepts <Gut> and <Brain> are widely represented in the biomedical terminological resource landscape on their own, our main focus at this stage, given the scope of the project, was to gather information about whether both concepts could be found in an explicitly interconnected way in the aforementioned resources. The results showed only 3 matches displaying this relationship (Interlinking Ontology for Biological Concepts (IOBC)¹³, Medical Subject Headings (MeSH)¹⁴, and Ontology for Host-Microbe Interactions (OHMI)¹⁵).

From a conceptual perspective, the concept under analysis seems to be categorized as either <Phenomenon> or <Process>, depending on the resource. Within the MeSH taxonomy, it is a type of <Nervous System Physiological Phenomena>, which is a subordinate concept of <Musculoskeletal and Neural Physiological Phenomena> and the latter, in turn, is a subordinate concept of <Phenomena and Processes Category> as a top-level concept. OHMI, on the other hand, is BFO-based¹⁶, with the concept under study being categorized as a type of <Host-Microbiome Interaction>, the latter being a subtype of <Interaction>, then <Process>, followed by <Occurrent>, and finally <Entity>. In IOBC, it is a subordinate concept of <Digestive System Physiology> which, in turn, is a subordinate concept of <Biological Phenomenon, Process, and State>, with <Phenomena> and <Terms Related to Life Science> as parent concepts.

At the linguistic level, three designations have been identified as preferred name/label: *brain-gut correlation*, *Brain-Gut Axis*, and *microbiome-gut-brain interaction*, for IOBC, MeSH, and OHMI, respectively. In OHMI, *microbiome-gut-brain axis* appears as a synonym, whereas in MeSH, a list of 25 synonyms is put forward. Although all these expressions represent the same concept in the MeSH database and have been created to serve the resource's indexing purposes, many of them do not follow term formation patterns in English and are thus hardly found in either oral or written discourse (e.g. Axis, Microbiota-Gut-Brain). Other expressions included in the MeSH list, however, might be useful to support corpus collection and processing within T3.3 (e.g. Gut-Brain Axis, Microbiota-Brain-Gut Axis, Brain-Gut-Microbiome Axis).

¹¹ <https://www.snomed.org/>

¹² <https://confluence.ihtsdotools.org/display/DOCGLOSS/postcoordinated+expression>

¹³ <https://bioportal.bioontology.org/ontologies/IOBC>

¹⁴ <https://bioportal.bioontology.org/ontologies/MESH> and <https://www.ncbi.nlm.nih.gov/mesh/>

¹⁵ <https://bioportal.bioontology.org/ontologies/OHMI>

¹⁶ Basic Formal Ontology (BFO) is the upper-level ontology upon which OBO Foundry ontologies are built (as is the case of OHMI). Cf. <https://obofoundry.org/> for more information.

Both MeSH and OHMI contain textual definitions, with expressions such as “interactive network” (“between the gastrointestinal tract (gut) and the brain”) and “interaction” (“between enteric microbiota on the host and the host brain”) being used to verbally represent the concept. Given the scarcity of linguistic data concerning <Gut-Brain Axis> amidst such concept-oriented resources, and the fact that the systematic study of natural language definitions is also one of the objectives of T3.3, additional searches were conducted among glossaries and specialized lexicographic resources.

In the *Glossary of the International Foundation for Gastrointestinal Disorders*¹⁷, “brain-gut axis” is the designation used, with the definition pointing towards a “continuous bi-directional flow of information and feedback” (“that takes place between the gastrointestinal tract, and the brain and spinal cord (which together comprise the central nervous system”). The other result was found in a glossary included in a paper by Codagnone et al. (2019), which uses the term “gut-brain axis”, and defines the concept as a “multidirectional biological system” (“comprising the central nervous system, the neuroendocrine and the neuroimmune systems, the gastrointestinal tract and components of the enteric and autonomous nervous system”).

Even though many of the ontologies do not contain exact matches to the target concept <Gut-brain>, plenty of the closely related concepts are present: <Enteric Nervous System> (36 ontologies), <CORTICOTROPIN-RELEASING FACTOR> (32 ontologies), <Vagus Nerve> (59 ontologies), and <Neuroendocrine Hormone> (5 related concepts in NCIT)¹⁸.

As an extremely valuable resource of Knowledge-Rich Contexts - KRCs (Meyer, 2001), these textual definitions - as well as those collected about the remaining concepts - will undergo a more detailed analysis in subsequent stages of T3.3. In what concerns the designation level, current data, though scarce, seems to point towards a possible variation that will be addressed in future work within this task: are “gut-brain axis” and “brain-gut axis” indeed terms? If so, which one is the most frequent? Would “microbiome” and “microbiota” be part of the full designation? And do these terms in fact designate the same concept?

3.2.2 <Neurodegenerative Diseases>

As mentioned earlier, one of the three disease groups of the HEREDITARY project has also been selected as a case study, to ascertain how (conceptual and linguistic) knowledge about such diseases is represented and organized in current biomedical resources. Four specific diseases encompass this group: <Amyotrophic Lateral Sclerosis>, <Frontotemporal Dementia>, <Multiple Sclerosis>, and <Parkinson’s Disease>. It should be noted that the results of the BioPortal and HeTOP searches presented here do not include resources in which a given concept entry is not directly related to the concepts under study or has been reused. To further elicit a more detailed conceptual representation, encompassing not only hierarchical but also non-hierarchical

¹⁷ <https://iffgd.org/resources/medical-definitions-glossary-dictionary/>

¹⁸ One of the recognized disorders related to the gut-brain interplay is Irritable Bowel Syndrome (IBS) (available in the Ontology of Consumer Health Vocabulary and SNOMED), which will be explored further in the project, along with the remaining diseases.

conceptual relations, all four concepts were also searched for in the SNOMED CT browser (International Edition)¹⁹.

To illustrate our point, we present and describe a single example.

3.2.2.1 <Amyotrophic Lateral Sclerosis>

This concept had 40 matches in both BioPortal and HeTOP, yet one of the resources was not working. Its categorization as a <Neurodegenerative Disease> of the <Central Nervous System> is confirmed by the available taxonomies, along with the <Motor Neuron> impairment of this pathology. The SNOMED CT concept diagram, with a status of Primitive, reflects this conceptualization, with <Amyotrophic Lateral Sclerosis (disorder)>²⁰ as a type of <Motor Neuron Disease> located in the <Structure of Nervous System (body structure)> (Annex 1).

In what concerns the linguistic dimension, the preferred designation is *amyotrophic lateral sclerosis* (with upper/lower case variation), with some of the proposed synonyms including the abbreviated form *ALS*, *amyotrophic lateral sclerosis with dementia*, *amyotrophic sclerosis*, *myelopathic muscular atrophy*, as well as the eponyms *Lou Gehrig disease* (also *Lou Gehrig's disease*, *Lou Gehrigs disease*, *Gehrig's disease*), *Charcot disease* (also *Charcot syndrome*, *Charcot's syndrome*), and *Aran-Duchenne disease* (also *Aran-Duchenne muscular atrophy*).

In the 39 entries, there were 17 textual definitions (11 unique, with MeSH's and the Disease Ontology's - DOID - definitions being reused), encompassing KRCs which may point towards certain essential characteristics of the concept (cf. Section 2): "is a neurodegenerative disease"/"disorder"; "is a motor neuron disease"; "is a progressive, fatal, neurodegenerative disease"; "is a nervous system disease".

3.3 Discussion and next steps

The results outlined in this section of the deliverable, albeit requiring a more in-depth analysis at later stages of the project, constitute an important first step to inform and support subsequent work within WP3 (especially in T3.1 and T3.3). Firstly, the association of <Gut> and <Brain> into a unique concept that reflects its interplay is still relatively underrepresented in biomedical terminological resources. Moreover, potential variation phenomena in current verbal designations seem to reflect the fact that research in this area is rather recent and knowledge is yet to be stabilized. Future work will look particularly into the *gut-brain vs. brain-gut* examples in the corpus, to elicit information that may help ascertain if, indeed, such designations are synonyms from a terminological perspective, i.e. if they designate the same concept.

The data collected from the use case on <Neurodegenerative Diseases> also provides useful reflection points: on the one hand, the conceptualizations appear to be relatively stable across resources (despite their differing functions and structural principles), yet it is relevant to point out that none of the diseases under analysis are Fully Defined

¹⁹ <https://browser.ihtsdotools.org>

²⁰ The semantic tag (disorder) is included in parentheses at the end following SNOMED CT editorial guidelines regarding the creation of Fully Specified Names (cf. <https://confluence.ihtsdotools.org/display/DOCEG/Fully+Specified+Name>).

concepts in SNOMED CT. Further work is needed to see if the same occurs with the remaining HEREDITARY diseases. Moreover, several URI links in some of the resources available via BioPortal were broken, hampering data reuse and making mapping endeavors into and from these resources particularly challenging.

On the other hand, the few collected textual definitions, although, in most cases, lacking a source, elicited partial information that may point towards the most relevant characteristics of each concept. The expansion of our definition subcorpus to include the remaining diseases, via not only the biomedical terminological datasets mentioned here but also other terminological and lexicographic resources, will allow ongoing work on textual definition analysis by several team members (cf. Carvalho et al., 2023; Bonato et al., 2024) to provide further insights. Other relevant linguistic expressions found in several definitions of the different concepts under study seem to be consistently used in certain categories, such as **signs and symptoms** (“clinical manifestations include...”; “signs and symptoms include...”; “symptoms include...”; “is characterized by...”; “is marked by...”; “this is manifested with...”; “with symptoms such as...”), **causes** (“caused by...”), **affected body structures** (“in which... are affected”; “affecting...”; “primarily affecting...”; “that primarily affect...”; “involving...”; “predominantly involving...”; “may also be found in...”), as well as **consequences** (“resulting in...”; “it results in...”; “causing...”). This is yet another aspect that needs further analysis and comparison, mainly to see whether one can indeed categorize the use of such expressions, reuse them accordingly as part of definitional templates, and thus contribute to more automated and consistent drafting of natural language definitions, in line with ongoing work in this regard (cf. Carvalho et al., 2018).

Finally, in what concerns verbal designations, the analyzed data presented examples of forms that can be useful for the corpus-based tasks, supporting corpus expansion and analysis, especially in studying potential cases of variation. An important aspect to highlight is that in some cases, the designations suggested as synonyms in the consulted resources seem to point towards contradictory information in relation to what is depicted in the conceptual representations. From a terminological standpoint, especially when based on the premises put forward by the ISO 1087 and 704 standards, to be considered synonyms, two terms must designate the same concept (and not a superordinate or a subordinate one).

In addition to being the starting point for subsequent work in the project, these initial results substantiate the pertinence of the HEREDITARY project, as well as the opportunity to make a valuable contribution to this domain, at both the conceptual and linguistic levels, in the current healthcare landscape.

4 Domain-specific corpus: tools and methods

Building and using corpora for terminological data extraction, and its subsequent analysis and/or description in terminological resources—a series of endeavors that partially define the *terminological work* (ISO 1087:2019)—is a broadly used methodology throughout different scientific communities.

A corpus is generally understood as a collection of texts used as a sample of language despite the wide array of corpus typologies. In the present project, the corpus of analysis will be a domain-specific textual corpus, namely a collection of texts written by experts for experts within the biomedical sciences, for our core methodological goals aim at (i) identifying, (ii) extracting, (iii) systematizing and (iv) analyzing terminological data—domain knowledge, specialized information and knowledge-rich contexts (KRC) (Meyer, 2001), among others—, that will allow us to infer the experts' conceptualizations, beyond terms.

In this section, we outline the criteria for text eligibility and the underlying goals guiding the corpus construction. We will also detail the process of documenting the corpus architecture as a linguistic and terminological resource, along with the associated metadata. The following subsections will describe our methods for corpus development, while also providing an overview of the current preliminary stage of the reference corpus of analysis— HEREDITermCorpus.

4.1 Corpus compilation process: criteria, typology, and goals

4.1.1 Full papers: a paramount criterion for terminological data extraction

The first task in building a domain-specific corpus focused on Parkinson's Disease (PD) and Alzheimer's Disease (AD) will involve validating full papers that discuss on PD and AD, in association with the term “gut-brain axis”. Validation is based on a manual analysis of each text, which is considered eligible for inclusion in the corpus if it meets the following core criteria: (i) relevance to PD and AD, (ii) availability in the public domain (open access), and (iii) machine readability. The third criterion ties with Sketch Engine²¹ (SKE), the NLP used for compiling, annotating, and exploiting the corpus. Finally, but not the less important criterion, the preferable sources for text eligibility are scientific journals and reputable biomedical publishers.

The process of text validation will be documented in a database to ensure that original text-related information is preserved throughout the corpus compilation.

4.1.2 Subcorpus: a collection of abstracts

Abstracts addressing the gut-microbiota-mental health axis will be a secondary source of texts to be included in the corpus given their short length of text, which implies fewer terminological data to be captured such as contextual definitions and definitional contexts (Ramos, Costa, & Roche, 2019). As with the previous validation tasks, abstracts will be

²¹ <https://www.sketchengine.eu/>. Sketch Engine has an open-source version (<https://www.sketchengine.eu/nosketch-engine/>).

assessed based on their relevance to the project's scope. If the criteria for corpus inclusion are met, abstracts will be compiled as a subcorpus within the main corpus, i.e., the corpus comprising full papers, to which we refer to as 'reference corpus of analysis'.

4.1.3 Challenges and opportunities

Given the limited number of full papers on the topics of interest that we currently have in our possession, we will resort to the WebBootCaT²² technology to create text corpora from web pages, with the help of SKE, to capture additional texts to the existing collection of full papers (cf. Section 4.1.1). This feature allows us to find web pages efficiently based on a set of "seed words", i.e., a list of keywords previously parameterized in the software's search configurations. The validation of these texts follows pre-established criteria, as presented in Section 4.1.1.

4.2 Documentation: text types, criteria, and metadata

Documenting the process of corpora building is considered a best practice in Corpus Linguistics (Baker, Hardie, & McEnery, 2006). The primary goal is twofold: on one hand, to ensure the corpus's reusability by making the information clear and accessible for new users and projects, and on the other hand, to support the interpretation of results during corpus exploitation. Metadata not only enables terminologists to quickly and accurately access relevant information during data analysis but also serves to (i) expedite the process of locating specific information within the corpus, or (ii) focus the analysis on particular metadata types, such as the text's topic.

4.2.1 Text-related and authorship-ID information

The validation of the collection of texts will be documented in a database with metadata assigned to each text to ensure that important text-related and authorship information is preserved. This includes details such as the PMID, title, publication date, source of text capture (which may not always align with the PMID), and DOI. In addition to this text-ID information, other attributes, such as text typology (e.g., Review, Protocol, etc.), were also recorded.

4.2.2 Statistical data of the domain-specific corpus for HEREDITARY

The corpus is currently under construction with Sketch Engine and designed to be divided into two parts: (1) a subcorpus composed of abstracts and (2) a collection of full papers. Whereas the subcorpus comprises 704 texts with 386,953 *tokens*²³, including words and non-words (e.g., alphanumeric sequences; punctuation), the collection of full papers consists of 293 texts with 4,751,426 tokens. Assembling the two parts, the main corpus serves as the reference corpus for analysis, comprising 997 texts with a total of 5,137,643 tokens. Of these tokens, 168,075 are unique forms (sequences of characters) also known as 'types', which occur a total of 3,512,651 times. The unique forms include alphanumeric and numeric sequences.

²² <https://euralex.org/publications/webbootcat-a-web-tool-for-instant-corpora/>

²³ An "individual occurrence of a *type* (3.29) in a *text corpus* (3.25)" (**Error! Reference source not found.**, 2024, p.4).

4.3 Corpus management and future directions

At this early stage, the corpus – HEREDITermCorpus – will be monolingual, but future plans include expanding it to a multilingual framework. The goal is to analyze terminological data at the conceptual level, without overlapping it with the (multi)linguistic dimension, as language systems are not isomorphic at the morphosyntactic level.

The corpus will be continually expanded with additional texts, requiring ongoing management of the corpus-building process documentation. Consequently, the design of HEREDITermCorpus will align with the typology of a multilingual, domain-specific, monitor corpus. Additionally, as the metadata in the database used for text systematization is based on pre-established criteria, the model will provide guidelines for the ongoing corpus compilation and related documentation.

4.3.1 Project-related data sources: legal and ethical aspects

The project-related data sources are (could be) key contributors to the corpus. However, legal and ethical considerations are critical when using patient-clinical data in any project. Compiling such sensitive information into a corpus is no exception. While traditional texts are read sequentially from top to bottom, a corpus—comprising a collection of texts—is queried as a whole, with the goal of identifying statistical data or specific linguistic or morphosyntactic patterns that allow the analyst to infer concept- or term-related information. Nonetheless, there are methods for anonymizing patient-clinical texts to remove personal details such as names, birth dates, and geographic locations. These techniques can be further explored with the designated partners involved in this project, to align our methods with the General Data Protection Regulation (GDPR)—a regulatory standard specifically focused on issues related to data privacy (cf. Regulation (EU) 2016/679, in EUR-Lex²⁴).

²⁴ <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

5 Terminology extraction

Computational Terminology (CT) stands at the crossroads of several disciplines, bringing together computer scientists, information specialists, linguists, and terminologists to design and implement automated methods for processing specialized texts. A milestone for the discipline came in 2001 with the publication of the first comprehensive volume devoted entirely to CT (Bourigault et al., 2001). This collection of essays explored a wide range of topics, from automated text parsing and terminology management to issues of information retrieval and multilingual database alignment.

In particular, the study of Automatic Term Extraction (ATE) was central, offering insights into its role in solving practical challenges such as translation and knowledge organization.

The origins of systematic research into ATE can be traced to the mid-1990s. For instance, Kageura and Umino (1996) provided a comprehensive review of early term indexing techniques dating back to the 1950s, starting with Luhn's pioneering work (1957). Their analysis also underscored key milestones in information retrieval, such as Sparck Jones's concept of term specificity (1972), and proposed criteria for defining a term and its termhood, including frequency within a domain, exclusivity to a domain, and relative prominence in a domain compared to general contexts.

Subsequent works on CT (Drouin et al., 2015, 2018) further highlighted emerging trends, particularly the integration of hybrid methods combining artificial neural networks and distributional semantics approaches, as exemplified by word embeddings based on the distributional hypothesis (Mikolov et al., 2013).

Building on those foundational works, this section focuses on recent advancements in ATE. Given the space limitation and the main objective of this Section, we do not intend to perform a systematic review of all the latest publications, rather we examine key developments and future directions, emphasizing the strategic importance of ATE as both a multidisciplinary bridge and a crucial tool for constructing multilingual terminological databases.

Our analysis started from the two most recent surveys for this field (Hanh, et al., 2023) and (Di Nunzio et al., 2023). In this deliverable, we have chosen to focus mainly on tools for automated text extraction that are available as open-source solutions online. This decision reflects our commitment to accessibility, transparency, and reproducibility in research and practical applications. This choice aligns with the broader objective of promoting equitable and open-source access to technological advancements.

As a complementary approach to the CT, we will tackle the linguistic approach for terminology work, where we will explore the methods and ongoing efforts related to semiautomated terminology extraction (ISO/FDIS 5078:2024) from the domain-specific corpus being developed for the project (cf. Section 4). The methodology for terminological data extraction through semi-automated methods, follows a 'hybrid leading criterion' (cf. ISO/FDIS 5078:2024), as it incorporates a combination of techniques and technologies, as further explained in Section 5.5.

5.1 Methodologies and approaches

The core methodologies and algorithms that power automated term extraction, ranging from traditional rule-based and statistical techniques to neural network and large language models approaches. We can roughly divide these approaches into the following categories: Linguistic Approaches, Statistical Approaches, Machine Learning approaches, Deep Learning Approaches:

- Linguistic approaches rely on the structure and rules of language to identify terms. In particular, Part-of-Speech Tagging (POS) and pattern-based methods identify noun phrases and multi-word expressions or use predefined syntactic patterns (e.g., adjective + noun or noun + noun) to extract candidate terms. Morphological tools too play a main role in analyzing and understanding the structure of words for term extraction. For example, lemmatizers which reduce words to their base or canonical form (lemma) and capture variants (e.g., plural vs. singular, derivational forms) of candidate terms.
- Statistical methods are data-driven approaches and focus on identifying term candidates based on corpus frequency and statistical properties. A few examples of these measures are: Term Frequency-Inverse Document Frequency (TF-IDF) which measures the importance of terms in a document relative to the entire corpus, Mutual Information (MI) which detects co-occurring word pairs or multi-word terms based on their likelihood of appearing together, C-value (and NC-value) which estimates the termhood of multi-word expressions based on their frequency and nested occurrences.
- Machine Learning-Based approaches are increasingly popular, but they require corpora for training, depending on the approach these corpora may be annotated or not (supervised or unsupervised learning). With these approaches, documents as well as terms are represented by features, which can include syntactic, semantic, and statistical features, and these text representations are the foundation for the models that need to be trained.
- Deep Learning approaches, specifically large language models, are a subset of machine-learning approaches but they are currently one of the most used and analyzed approaches which can process contextual information for sophisticated term identification. In this context, term embeddings is a mathematical representation of terms as dense, continuous vectors in a high-dimensional space, capturing their semantic and contextual relationships. This semantic representation enables ATE systems to identify terms by analyzing their context-aware relationships, disambiguating polysemous terms, and grouping related variations. Embeddings significantly enhance the ability to extract domain-specific terms by understanding their underlying semantics rather than just surface patterns.

5.2 Tools and software for term extraction

This section reviews existing tools and software solutions for term extraction. It aims to provide readers with a practical understanding of the tools available and their potential applications. We present these tools according to the same categorization of Section 5.1 considering that, even in this case, the categorization is fuzzy and that tools can be classified under different labels given their features.

- Linguistic Approaches tools:
 - TreeTagger (<https://www.ims.uni-stuttgart.de/en/research/resources/tools/treetagger/>): For POS tagging and lemmatization, mainly used for term pattern identification.
 - Unitex/GramLab (<https://unitexgramlab.org/language-resources>): A corpus processing tool for pattern-based linguistic term extraction.
 - Text2Onto (<https://code.google.com/archive/p/text2onto/>): Combines linguistic rules with ontology generation.
 - Stanford NLP (<https://nlp.stanford.edu/>): Provides POS tagging and syntactic parsing to identify term candidates.
 - Python NLTK (<https://www.nltk.org/>): the Natural Language toolkit which provides a wide variety of solutions for POS.
 - MetaMap (<https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html>): Specialized for extracting medical terms using linguistic analysis and ontology matching.

- Statistical Approaches tools:
 - YAKE! (Yet Another Keyword Extractor) (<https://liaad.github.io/yake/>): Extracts keywords based on statistical features like word frequency and contextual metrics.
 - RAKE (Rapid Automatic Keyword Extraction) (<https://csurfer.github.io/rake-nltk/>): Identifies multi-word terms based on word co-occurrence and positional weighting.
 - AntConc (<https://www.laurenceanthony.net/software/antconc/>): A corpus analysis tool that uses frequency-based methods for term extraction.

- Machine Learning Approaches:
 - Weka (<https://ml.cms.waikato.ac.nz/weka/index.html>): A general machine learning toolkit often used for training classifiers for term extraction.
 - Python Scikit-learn (<https://scikit-learn.org/>): Widely used for custom-built supervised or unsupervised ATE pipelines.
 - Gensim (<https://pypi.org/project/gensim/>): For unsupervised topic modeling (e.g., Latent Dirichlet Allocation) to identify domain-specific terms.
 - OpenNLP (<https://opennlp.apache.org>): Provides machine learning models for NLP tasks, including term and named entity extraction.

- Large Language Models (LLMs):
 - Python SpaCy (<https://spacy.io/>): Combines SpaCy's linguistic pipelines with transformer models for accurate term extraction.
 - BERT (Bidirectional Encoder Representations from Transformers) (<https://github.com/google-research/bert>): Fine-tuned for identifying terms and performing keyphrase extraction in domain-specific corpora.
 - Hugging Face Transformers (<https://huggingface.co/>): A versatile library supporting LLMs like BERT, RoBERTa, and GPT for ATE tasks.

- GPT models (https://github.com/GPT-Alternatives/gpt_alternatives): Used for extracting terms by generating or understanding text with contextual semantics.
- BioBERT (<https://github.com/dmis-lab/biobert>) and SciBERT (<https://github.com/allenai/scibert>): Specialized for biomedical and scientific term extraction.

5.3 Evaluation metrics and benchmarking

The survey compiled by (Hanh et al, 2023) provides the most comprehensive list of metrics and benchmarks for evaluating Automatic Term Extraction (ATE) tools. In this section, we will summarize the most important points and leave the reader the possibility to explore this topic further.

Since ATE is basically a labelling problem – in its simplest form, the ATE tool has to select the portion of the text that represent a candidate term – the key evaluation metrics of this task are related to those of text classification and retrieval:

- Precision, Recall, and F1 Score: These are core metrics used to measure the performance of ATE tools. Precision assesses the proportion of correctly identified terms among the extracted terms, while Recall measures the proportion of gold-standard terms identified. F1 Score combines these into a harmonic mean, offering a balanced view.
- Ranking Metrics: ATE tools often generate ranked lists of term candidates. Metrics like Average Precision (AvP) are used to evaluate the quality of these rankings, providing insights into how well the top-ranked terms match the gold standard.
- Corpus-Level vs. Document-Level Evaluation: Evaluation can occur at Corpus-Level, if one considers the entire dataset to determine overall performance, or at Document-Level, if one focuses on individual documents, measuring the tool's consistency and adaptability across different contexts.

These metrics are fundamental for comparing tools across domains and languages. However, in order to compute these measures and compare the different ATE approaches, we need standard benchmarks. Some of the benchmarks are domain-specific, often favoring tools tailored to specific types of corpora, such as biomedical datasets, others are more oriented towards general domains. Once again, we summarize some of these benchmarks referring to the survey:

- The ACTER (Annotated Corpora for Term Extraction Research) dataset (<https://github.com/AylaRT/ACTER>) is one of the most important multilingual resources designed to benchmark ATE systems. It spans four specialized domains - Corrosion Engineering, Wind Energy, Heart Failure, and Orchard Management - and supports English, French, Dutch, and German. The dataset includes gold-standard annotations for Single-Word Terms (SWTs) and Multi-Word Terms (MWTs), with annotations standardized across domains and languages. Its focus on domain-specific, multilingual term evaluation and public availability makes it a critical resource for comparing linguistic, statistical, and neural methods. While its coverage

- may not extend to broader domains like biomedicine, it has been used in shared tasks to standardize ATE evaluation and test cross-domain and multilingual systems.
- The ACL Anthology Reference Corpus (ACL ARC) (<https://github.com/languagerecipes/acl-rd-tec-2.0?tab=readme-ov-file>), derived from computational linguistics research papers, is valuable for ATE evaluation in technical writing. The ACL RD-TEC 2.0 has been developed with the aim of providing a benchmark for the evaluation of term and entity recognition tasks based on specialized text from the computational linguistics domain. This release of the corpus consists of 300 abstracts from articles in the ACL Anthology Reference Corpus, published between 1978–2006. In these abstracts, terms (i.e., single or multi-word lexical units with a specialized meaning) are manually annotated.
 - The GENIA corpus (<https://paperswithcode.com/dataset/genia>), focused on biomedical terminology, is extensively annotated with terms such as proteins and genes, making it an ideal resource for evaluating biomedical text-mining tools. The GENIA corpus is derived from MEDLINE abstracts related to topics like transcription factors, proteins, and cellular processes. Its focus on detailed biological entities makes it particularly suitable for tasks such as named entity recognition (NER), term extraction, and relation extraction within biomedical literature.
 - The SimpleText Lab at CLEF 2024 focuses on making scientific information accessible to a broad audience (<https://simpletext-project.com/2024/en/>). In particular, Task 2, "Identifying and Explaining Difficult Concepts," is especially relevant for term extraction. Participants are tasked with identifying up to five difficult terms from scientific abstracts and providing clear definitions or explanations for each. The dataset for this task is derived from scientific abstracts primarily in computer science and engineering. It includes training, validation, and test sets annotated with terms, their associated difficulty levels (easy, medium, or difficult), and intentional definitions. This dataset is especially valuable for testing systems aimed at handling complex terminology and generating user-friendly explanations.

5.4 Specialized term extraction: medical domain

In the previous sections, we presented a general overview of the main approaches, tools, and benchmarks for ATE. In this section, we focus ATE in the medical domain since there are some specialized methods, tools, and benchmarks to address the unique challenges posed by the complexity and specificity of medical language.

We have already presented MetaMap in Section 5.2 as one of the available linguistic tools for ATE. More specifically, MetaMap leverages domain-specific rules by means of an analysis of morphological features such as prefixes, suffixes, and stems and ontologies, like UMLS, to identify medical terms.

The cTAKES (clinical Text Analysis and Knowledge Extraction System) toolkit (<https://ctakes.apache.org/>) is an open-source (NLP) tool specifically developed for processing clinical narratives. It has been widely used to extract information such as medical concepts, events, and relationships from unstructured clinical texts like electronic health records (EHRs). A key feature of cTAKES is its integration with

biomedical knowledge bases such as the UMLS, allowing it to identify and map entities like diseases, drugs, symptoms, and anatomical terms to standardized ontologies.

SciSpacy (<https://allenai.github.io/scispacy/>) is an efficient NLP toolkit specifically designed for scientific and biomedical text. Built as an extension of SpaCy, see Section 5.2, SciSpacy incorporates pre-trained models and pipelines that are fine-tuned for extracting biomedical entities and terms. It provides access to a range of vocabularies and ontologies, such as UMLS, MeSH, and SNOMED CT (cf. Section 3 of this deliverable), allowing for comprehensive coverage of biomedical and clinical terminologies. Key features of SciSpacy include named entity recognition (NER) for identifying diseases, drugs, genes, and other entities, as well as entity linking to map extracted terms to their corresponding entries in biomedical ontologies.

Deep learning models introduce advanced capabilities through pretrained embeddings like BioWordVec, BioBERT, SciBERT, and ClinicalBERT which capture semantic and contextual nuances in biomedical literature.

BioWordVec (<https://github.com/ncbi-nlp/BioWordVec>), based on Word2Vec, focuses on creating word embeddings for biomedical text by training on large-scale datasets such as PubMed and MIMIC-III. It captures semantic relationships between words, making it useful for simpler tasks like clustering, similarity analysis, and feature extraction in biomedical research. On the other hand, BioBERT (<https://github.com/dmis-lab/biobert>) extends the popular BERT model by pretraining on biomedical corpora, such as PubMed abstracts and PMC full-text articles. This specialized training equips BioBERT to handle complex tasks like named entity recognition, relation extraction, and question-answering with significantly improved accuracy compared to general-purpose language models. Similarly, SciBERT (<https://github.com/allenai/scibert>), another adaptation of BERT, is tailored for scientific literature across multiple disciplines, including biomedicine, by pretraining on a corpus of more than one million scientific papers from Semantic Scholar.

ClinicalBERT (<https://github.com/kexinhuang12345/clinicalBERT>) is another specialized variant of the BERT model designed specifically for understanding clinical language, particularly text derived from electronic health records EHRs. By fine-tuning the original BERT model on clinical datasets such as MIMIC-III, which contains de-identified patient notes and discharge summaries, ClinicalBERT captures the nuances and terminology unique to clinical narratives.

Regarding the datasets, in addition to the GENIA dataset already presented in Section 5.3, there are other useful dataset for the training and evaluation of ATE tools in the medical domain.

For example, the BioCreative datasets (<https://paperswithcode.com/dataset/bc5cdr>) developed as part of the BioCreative challenges support shared tasks in biomedical natural language processing, such as named entity recognition, term normalization, and relation extraction. They cover a range of domains, including drug-gene interactions and chemical-protein relationships.

The MIMIC-III (<https://physionet.org/content/mimiciii/1.4/>) is a large dataset of de-identified electronic health records from real patients. MIMIC-III contains clinical notes, discharge summaries, and other medical documents annotated for terms related to diseases, symptoms, treatments, and procedures. It is a valuable resource for clinical text mining and predictive modelling.

The SemEval Biomedical Tasks (<https://semeval.github.io/SemEval2024/tasks.html>), as part of the Semantic Evaluation (SemEval) challenges, focus on extracting biomedical terms, identifying semantic relationships, and classifying medical concepts. These datasets provide a framework for comparing ATE systems on semantic and contextual accuracy.

The i2b2 Clinical NLP Challenges Datasets (<https://paperswithcode.com/dataset/2010-i2b2-va>) were developed as part of the i2b2 challenges and contain annotated clinical notes for tasks such as medical term extraction, relationship identification, and temporal reasoning.

5.5 Semi-automated terminology extraction: tools and methods

This subsection focuses on the ongoing efforts regarding terminological data extraction from the domain-specific corpus being compiled within the HEREDITARY project.

For the compilation, annotation and corpora exploitation, we use Sketch Engine²⁵, a natural language processing (NLP) tool. Our option ties with the several embedded tools, namely the corpus annotation (automatic and manual), along with the text type analysis via the statistics of the manually assigned metadata to each text. The added value of the annotation, together with appropriate tools for corpus analysis, offers the terminologist a wide range of approaches to the corpus. The terminologist can extract data (language evidence) from the corpus by means of specific queries, where lemma²⁶, Part-of-Speech (POS), or morphosyntactic structures are parameterized with the help of artificial languages used in computer science, such as *regular expressions*, also known as

²⁵ Sketch Engine combines hybrid corpus compilation (users' texts and WEB corpora), automatic annotation of POS for CQL (corpus query language) queries with *regex* (regular expressions), manual annotation of metadata, and terminological data extraction via text-type analysis through the statistics of their metadata, just to mention the core elements used for the present project.

²⁶ According to **Error! Reference source not found.** et al., it is the "[t]he canonical form of a word (the correct Greek plural is lemmata, although some people write the plural as lemmas and may consider lemmata to be somewhat pedantic). [...] Lemmatized forms are sometimes written as small capitals, for example the verb lemma walk consists of the words walk, walked, walking and walks. In corpus studies, word frequencies are sometimes calculated on lemmata rather than types; words can also be given a form of annotation known as lemmatization. (2006, pp. 103-104).

*regex*²⁷ – a feature commonly used in corpus query languages²⁸ (CQL) with NLP tools, such as Sketch Engine (SKE). The purpose of using regexes is to find patterns, such as frequently co-occurring lexical units. In our view, these patterns are usually indicators of specialized knowledge, e.g., terms and lexical-semantic relations, to the extent that we prefer this method at a later stage in the approach to the corpus.

The semi-automated terminology extraction approach employed in this work follows a hybrid leading criterion (cf. ISO/FDIS 5078:2024), as it incorporates a combination of (a) techniques and (b) technologies:

- (a) Statistical – frequency, termhood²⁹ and association:
 - Linguistic – POS (Part-of-speech), used for queries:
 - (i) to match linguistic data that falls under a given grammatical category,
 - (ii) aiming at morphosyntactic patterns, (e.g., predicate → object, via Word Sketch³⁰),
 - (iii) using degrees of association between lexical units– *unithood*³¹– to capture lexical units that frequently co-occur (e.g. n-grams).
 - (b) Rule-based – using:
 - (i) formal patterns in CLQ syntax,
 - (ii) meta-information annotated to each text in the pre-processing stage of the corpus compilation, and
 - (iii) POS patterns to match, for instance, lexical markers pointing at knowledge patterns.
- (c) The combination of (a) and (b).

5.5.1 Terminological data systematization

The quantitative results of the corpus exploration will be systematized in a database, complemented by graphical representations to visualize the distribution of each form within the corpus. Observing whether a form appears in one or multiple texts provides insights beyond the raw statistics, particularly indicating that the widespread use of a

²⁷ A regular expression is a sequence of characters that forms a search pattern. When you search for data in a text, you can use this search pattern to describe what you are searching for. A regular expression can be a single character, or a more complicated pattern. Regular expressions can be used to perform all types of text search and text replace operations. https://www.w3schools.com/js/js_regex.asp

²⁸ According to the SKE terminology, “[t]he Corpus Query Language is a special code or query language used in Sketch Engine to search for complex grammatical or lexical patterns or to use search criteria which cannot be set using the standard user interface.” <https://www.sketchengine.eu/documentation/corpus-querying/>

²⁹ According to **Error! Reference source not found.**: “degree to which a lexical unit (3.8) is recognized as a term (3.19)” (2024, p. 3).

³⁰ “[A] one-page summary of the word’s grammatical and collocational behavior.” <https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/>

³¹ According to the ISO TC37/SC3: “degree to which a given sequence of words has sufficient collocational strength to form a stable lexical unit” (**Error! Reference source not found.**, 2024, p. 4).

form by experts strongly suggests it is a term. Additionally, examples of contexts obtained from the concordances are meant to be included in this data systematization.

5.5.2 Data analysis

The terminological data analysis will ground on a *mixed approach* (Costa, 2006), i.e. either from the conceptual or the linguistic level, depending on the meaning that relevant (not necessarily statistical) linguistic data encapsulate, or on the concept-related information conveyed by a given context. The regularities or singularities observed in texts are strong indicators of specialized knowledge information. The analyses of the latter will be conducted using the methods pointed in Section 5.5.3 and Section 5.5.4.

5.5.3 Lexical markers pointing at lexical-semantic relations

Lexical markers are linguistic expressions that commonly point at lexical-semantic relations with a prime terminological goal: they provide us with coordinates that guide us through the task of organizing knowledge information (Ramos, 2020). Many authors refer to this common feature as knowledge patterns found in knowledge-rich contexts (KRC) and state that some of these patterns are context dependent if we think of domain-specific fields of interest (cf. Meyer, 2001; L'Homme, 2004; Marshman, L'Homme, & Surtees, n.a.).

In our perspective, the linguistic analysis of lexical markers found in co-text with terms is paramount for modeling specific domain knowledge. Specialized texts do not convey all the necessary information that enables non-experts to grasp the experts' conceptualizations, therefore the need from the terminologist for an in-depth linguistic analysis of the morphosyntactic behavior of all lexical items present in a given definitional context.

Building on the assumption that relationships are fundamental to concept creation, and that "creating meaning through language also requires analyzing the relations between words in a sentence (or text)" (cf. Lim, Liu, & Lee, 2011), the terminologist can only address the conceptual dimension after conducting a linguistic analysis of the terminological data, as the conceptual aspect relies on the linguistic one. Therefore, the methods employed in this phase of the terminological work involve several distinct tasks:

- (1) a linguistic analysis of contexts and contextual definitions is carried out for the identification of specialized information, stemmed from the linguistic expressions in co-text with terms,
- (2) the systematization of the lexical-semantic relations pointed at by the lexical markers found in contexts and definitional contexts,
- (3) based on the systematization in (2), modeling the linguistic information in the form of lexical maps,
- (4) based on the lexical-semantic relations systematized in (2), the identification and systematization of the corresponding conceptual relations,
- (5) based on the systematization in (4), modeling the conceptual relations in the form of conceptual maps, to propose a micro-conceptual system of the domain under focus.

5.5.4 Rule-based methods for lexical-semantic relations identification

At this stage, efforts have already been made to identify and extract linguistic expressions that indicate lexical-semantic relations between terms from the HEREDITermCorpus. These linguistic expressions are access points to knowledge patterns, which we refer to as lexical markers (LM) (Ramos, 2020; Ramos & Costa, 2024).

The subject on knowledge patterns that suggest lexical-semantic relations between terms extracted from KRC's has been widely debated in literature (cf. Winston, Chaffin, & Hermann, 1987; Condamines & Rebeyrolle, 2001; Cruse, 2002; Barrière, 2004; Marshman, 2007; Johansson, 2008; Halskov & Barrière, 2010; Ramos, 2020; Ramos & Costa, 2024). As corroborated by Marshman (2010), knowledge patterns are commonly used to infer hierarchical conceptual relations such as generic–specific and part-whole, and non-hierarchical ones, such as the associative and cause–effect relations. Given the broad pragmatic scope of these two closely related relationships, a classification of subtypes urged to be identified, and with particular interest in the domain of medicine, as developed by Barrière (2002), cited by (Marshman, 2010), for the associative subtypes, and by Feliu (2004) and Nuopponen (2005), cited by (Marshman, 2010), for the subtypes of causal relationships.

As a short example for this work, we will focus on one sub-type of causal relationship: INCREASE (causing some characteristic of an entity or event to become “more”) and on one sub-type of associative relationship: CORRELATION (cf. Marshman, 2010).

Bearing in mind that cause–effect relations are commonly expressed by knowledge patterns such as *X causes Y*, we will use rule-based methods to capture LMs within two variables, namely (X, Y), which are frequently terms denoting concepts. Experiments with such rules (CQL) demonstrate that LM's pointing at associative or causal relationships are linguistically expressed through different forms, yet semantically denoting the same feature—e.g., falling under the subtype INCREASE, we can identify “provoke” and “contribute to” (cf. Annex 2).

The lexical-semantic analysis is not documented, for the work's stage is currently at terminological data extraction and systematization.

6 Conceptual and lexical relation typologies

6.1 Introduction to conceptual and lexical relationship typology

Terminology science is a discipline marked by the interconnection of the conceptual and the linguistic dimensions, that focus respectively on concepts and terms (Costa 2013; Santos and Costa 2015). Analyzing the conceptual dimension entails studying concepts and the relationships established among them, both of which are organized into conceptual systems. Concept relations are mirrored in the linguistic dimension of terminology through lexical relations. Lexical relations, represented in lexical networks, specifically concern relationships established among terms, thus being distinct from concept relations.

6.2 Typology of concept relationships

6.2.1 ISO 1087's approach to concept relationship typology

According to ISO 1087-1:2019, concept relations are classified into two main categories: hierarchical relations and associative relations. Associative relations are also defined as non-hierarchical concept relations.

6.2.1.1 Hierarchical relationships (e.g., is-a, part-of)

Hierarchical concept relations are categorized into two distinct types, namely generic relations (also called generic concept relations or genus-species relations) and partitive relations (also called part-whole relations or part-of relations).

A genus-species 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” (ISO 1087-1:2019). This type of relation is therefore established between a superordinate concept (generic concept) and a subordinate concept (specific concept). For instance, a hierarchic generic relation can be identified between <Vehicle> and <Car> (ISO 1087-1:2019), in the context of which <Vehicle> is the superordinate concept and <Car> is the subordinate concept. In conceptual systems, the relationship is graphically conveyed by the relation marker is-a.

A partitive relation, instead, is a “concept relation between a comprehensive concept and a partitive concept” (ISO 1087-1:2019). It is therefore a relation identified between a comprehensive concept “viewed as a whole consisting of various parts” and a partitive concept “viewed as a part of a whole”. Considering the example provided by the ISO 1087 (2019), this relation occurs between <Pedal> and <Bicycle>, as <Bicycle> constitutes the comprehensive concept and <Pedal> the partitive concept. In conceptual systems, the relationship is graphically represented by the relation marker is-part-of.

6.2.1.2 Associative relationships (e.g., cause-effect, temporal)

Associative relations encompass all types of concept relationships, with the exception of genus-species relations and partitive relations. By way of example, the relation that links the concepts <Education> and <Teaching> is an associative relation (ISO 1087-1:2019).

According to the categorization provided by ISO 1087 (2019), a type of associative relation is the sequential relation, specifically defined as an “associative relation by which concepts can be ordered by a relevant ordering criterion”. The ordering criteria are of a spatial, temporal or cause-effect nature. In particular, spatial relations are “based on the criterion of relative location in space”, therefore established between the concepts <Floor> and <Ceiling>. Temporal relations are “based on the criterion of following or preceding in time”, such as <Production> which is temporally prior to <Consumption>. Finally, causal relations (also called cause-effect relations) are “based on the criterion of cause and its effect”. A causal relation can be for example identified as existing between the concepts <Action> and <Reaction>.

6.2.2 Nuopponen’s approach to conceptual relationship typology

A different approach to conceptual relationship typology has been proposed by Anita Nuopponen (1994, 2005, 2022). In the work published in 2022, the author identifies 7 macro-groups of conceptual relations: generic relations, contiguity relations, activity relations, origination relations, developmental relations, interactional relations and causal relations. Each macro-group encompasses multiple concept relations.

- Generic relations are relationships established between superordinate concepts and subordinate concepts. These relations also comprise relationships that exist between coordinate concepts that are “subordinate concepts on the same level of abstraction”.
- Contiguity relations include various relations typologies: partitive relations, material-component relations, property relations, locative relations, enhancement relations, ownership relations, rank relations and temporal relations.
- Activity relations “are a set of concept relations, where one of the related concepts represents an activity”. These relations, for instance, link an activity to the “entity performing the activity (agent), the object of the activity, or the tools, materials or methods used”. Consequently, an agent relation exists between <Research> and <Researcher>.
- Origination relations are relations that “exist between concepts that refer to a concrete or abstract object and those that refer to its origin”. An example of these relations is the originator relation that exists between <Bread> and <Baker>.
- Developmental relations are “based on objects that go through stages in various types of process”. An example of developmental relations is the ontogenetic relation, that is established between <Child> and <Adult>.
- Interactional relations are “based on the interplay between the objects of reference”. For instance, a representational relation involves “an object and its representation”, thus linking <Term> and <Concept>, as well as <Place name> and <Place>.

Causal relations are relations that are not exclusively limited to the existence of a cause and an effect, because they also involve “temporal components”. Causal relations, for example, include explanatory causal relations, which is the kind of relation that exists between <Exposure to SARS-CoV> and <COVID-19>.

6.3 Typology of lexical relationships

The study of the linguistic dimension of terminology encompasses identifying the lexical relationships established among terms, including the relation between hypernyms and hyponyms and the relation between meronyms and holonyms.

6.3.1 Synonymy and near-synonymy

In terminology, synonymy refers to a relation between two terms in the same language that represent the same concept. From a discursive standpoint, however, two terms are considered synonyms if they can be used interchangeably, although there may be subtle differences in connotation, usage, or context. (Costa, 2017)

Near-synonymy refers to a relation between two terms that represent two concepts, where at least one characteristic distinguishes one concept from the other. In discourse, near-synonyms are closely related but may be used in different contexts or carry slight distinctions in meaning.

6.3.2 Hypernym and hyponym

As mentioned earlier, concept relations that are established in the conceptual dimension of terminology are mirrored in the linguistic dimension by way of lexical relations. The superordinate concept that pertains to the conceptual dimension may correspond to the hypernym at the linguistic level. The subordinate concept, instead, may correspond to the hyponym at the linguistic level. For example, the term “vehicle” constitutes the hypernym, whereas the term “car” is the hyponym.

6.3.3 Meronym and holonym

At the linguistic level, the lexical relationship between meronyms and holonyms can also be identified. In particular, the comprehensive concept that pertains to the conceptual dimension may correspond to the holonym at the linguistic level. The partitive concept, on the contrary, may correspond to the meronym at the linguistic level. It is therefore possible to identify a lexical relation between the terms “pedal” and “bicycle”, in the context of which “bicycle” constitutes the holonym and “pedal” is the meronym.

6.4 Conceptual and lexical relationship typology in the medical domain

Conceptual and lexical relationship typologies constitute a means to organize medical knowledge respectively concerning concepts and terms in medical resources. However, different typologies of relations are adopted in medical resources, depending on the specific categorization adopted by each resource.

6.4.1 Hierarchical and categorical relationships in UMLS

Medical knowledge is organized in the UMLS semantic network through the usage of 54 different types of relationships.³² These relationships include the “is-a relationship”. This

³² https://www.nlm.nih.gov/research/umls/new_users/online_learning/SEM_004.html

type of relationship establishes a conceptual connection between the concepts <Human> and <Mammal>, indicating that Human is-a Mammal.

The main non-hierarchical relationships are: `physically_related_to`, `spatially_related_to`, `temporally_related_to`, `functionally_related_to` and `conceptually_related_to`. However, these relationships, referred to as “semantic relationships”, do not always correspond to the relations established at the concept level.

In UMLS, the Parent-Child relationship (also referred to as the Broader-Narrower relationship) stands out as a significant relation that hierarchically links concepts within the biomedical domain. In this relationship, the Child represents a “subtype” of the Parent. For example, <Finding> is considered the Parent of <Sign or Symptom>, which is the Child.

6.4.2 Hierarchical and categorical relationships in SNOMED CT

The hierarchical concept relationship used in SNOMED CT³³ is the is-a relationship. For example, in SNOMED CT, <Infective pneumonia> is linked to <Pneumonia> through the relation marker is-a, which indicates that <Pneumonia> is a superordinate concept with respect to <Infective pneumonia>.

³³ <https://confluence.ihtsdotools.org/display/docstart/4.+snomed+ct+basics>

7 Health terminology validation

7.1 Key aspects promoting validation

The validation of terminology plays a crucial role in promoting clarity and accuracy in various health-related fields. Advances in health are significantly supported by accurate terminology, ensuring consistency in research, diagnosis, and treatment. Furthermore, effective communication in the health field depends, in part, on the use of validated terms that help increase communication between stakeholders, promoting better understanding and trust for all. On the other hand, increased health literacy is deeply linked to accessible and understandable terminology, enabling individuals to make informed decisions about their health through access to clear information. Taken together, these aspects underline the importance of sound terminological validation to improve the quality and accessibility of health.

7.1.1 Advances in health

Medical terminology constantly evolves to reflect advances in science, research, and technology, as well as changing public health priorities and the increasing personalization of treatments. This continuous evolution of science requires healthcare institutions to regularly review and update the terminology to ensure it reflects current practices and knowledge, supporting effective communication within and between healthcare disciplines (WHO, 2019). Investigating new diseases, treatment methods, and health conditions leads to the in-depth study of concepts, terms, and all relevant information that characterizes these conceptual and linguistic entities.

With the advancement of personalized and precision medicine, the complexity of knowledge generated in health is increasingly greater. Communication with patients requires intervention protocols to enhance individualized and efficient treatments. It is essential to help doctors and researchers accurately convey personalized approaches, which combine information of different natures, ranging from genetic and environmental factors to each patient's lifestyle. These advances in genomics and biotechnology have introduced numerous new terms linked to genetic markers and new therapies, which require systematization, description, or even standardization to ensure accurate understanding and application in clinical practice, minimizing the risks of misinterpretation (Ginsburg & Phillips, 2018).

Another important factor is the expansion of digital health technologies, including telemedicine, mobile health applications, and wearables, which have increased healthcare accessibility and generated vast data flows (Topol, 2019). Adopting validated terminology in this context facilitates seamless data exchange and ensures consistency across all digital platforms, allowing both patients and providers to interpret health information effectively (Frieden, 2017). Finally, the impact of AI and machine learning on healthcare underscores the need for standardized terms, as these systems rely on validated data to accurately interpret clinical information. This consistency increases the reliability of AI-based diagnoses and treatment recommendations, reducing the risk of

error in complex healthcare applications (Hersh, 2018). These elements highlight the importance of terminological validation in modern healthcare advances.

7.1.2 Improving health communication

Validating health terms brings significant benefits to health communication. First, improved accuracy and consistency in terminology reduce misunderstandings and ensure clear, standardized communication across healthcare settings, enhancing the reliability of shared information. The use of validated medical terminology tailored for different communication scenarios offers advantages by guaranteeing the precision, consistency, and reliability of exchanges.

The promotion of effective communication between doctors and patients involves adopting strategies such as active listening, empathy, and collaborative decision-making. The use of validated terms improves patient understanding and engagement, as clear and simple language helps patients grasp their diagnoses and treatment plans, empowering them to make informed decisions (Silverman, Kurtz, & Draper, 2013). Techniques such as open-ended questions, simplified language, non-verbal communication, the teach-back method, and encouraging patient questions build trust and help patients feel valued (American Medical Association, 2021; Institute for Healthcare Improvement, 2022). Furthermore, training health professionals in communication skills and emphasizing the importance of popularizing terms to promote health literacy (HL) are essential for patient-centered care (WHO, 2016).

Term validation also helps reduce medical errors by providing a common understanding that decreases the risk of miscommunication, particularly in high-stakes settings like emergency care (Kohn, Corrigan, & Donaldson, 2000). Additionally, trust and credibility are strengthened when terminology is clear and consistent, fostering public confidence in health information, especially during crises (Rector, Brandt, & Schneider, 2011). Validated terms are also crucial for multilingual and cross-cultural communication, ensuring that health messages remain accurate and meaningful across diverse languages and cultural contexts.

In the digital sphere, effective digital health communication and AI applications depend on validated terminology. This ensures reliable interactions with AI tools and accurate patient data collection, enhancing trust in these technologies (Topol, 2019). Finally, validated terminology enhances public health education and awareness by supporting clear messaging, clarifying complex health concepts, and combating misinformation, ultimately promoting positive health behaviors (Nutbeam, 2008). Moreover, it fosters a more inclusive and trustworthy digital health ecosystem by improving digital health literacy (DHL) and empowering users to make informed decisions about their health (Sørensen et al., 2012).

7.1.3 Increasing Health Literacy (HL)

The validation of health terminology plays a key role in improving HL, benefiting patients and citizens in general, particularly in terms of supporting the understanding of health-related messages, terms, and concepts. HL empowers people to manage health and

well-being in everyday life, it is recognized as a key determinant of public health that affects people's ability to make informed decisions and navigate health systems, as well as contributing to equity and efficiency in healthcare (WHO, 2013). “Health literacy is linked to literacy and entails people’s knowledge, motivation and competences to access, understand, appraise and apply health information to make judgments and take decisions in everyday life concerning healthcare, disease prevention, and health promotion to maintain or improve quality of life during the life course.” (Sørensen *et al.*, 2012). Therefore, by assessing and adapting to patients’ literacy levels, using plain language and visual aids, it is possible to improve patients' ability to understand their health information, which supports better adherence to medical advice.

Methodologies from linguistics and terminology, based on promoting the understanding of concepts and the appropriate use of terms in context, contribute to improving HL by strengthening personal competencies and abilities. This requires validated and consensual terminology, standardized where necessary, to ensure the accuracy of health information. Clear medical language reduces ambiguity, especially when complex concepts are explained in a simplified way so that as many people as possible can understand and use them (Silva *et al.*, 2023). Given the special emphasis on the empowerment and education of citizens that literacy promotes, the concept of <Lifelong learning> also encompasses the concept of <Digital health literacy> (DHL).

In a context where digital technologies are increasingly used to transmit health information and support clinical decisions, the provision of validated health terminology plays a crucial role in improving literacy, through its dissemination, and in empowering people to use digital resources. This approach refers to “the activation of competences and the demonstration of aptitude to carry out a set of actions that involve cognitive effort” and, at the same time, “the skills to use electronic devices and interact successfully with them” (Norman & Skinner, 2006). For example, validated terminologies like SNOMED CT³⁴ enable the creation of patient-friendly educational materials, bridging gaps in understanding and fostering active engagement in health management.

Validating terminology is not only important for non-experts in health, this approach is also useful for experts who are health professionals. For them, validated health terms improve communication between multidisciplinary teams, reducing ambiguities and errors in documentation and verbal interactions. From a digital point of view, a standardized language facilitates the exchange of data in electronic health records (EHRs), improving the accuracy of patient records and supporting clinical decision-making (Chute & Cohn, 2019). In addition, validated medical terms contribute to the transmission of consistent health messages, allowing professionals to align their instructions with patients' levels of understanding, thus improving adherence to treatment plans (Martins *et al.*, 2024).

By addressing both patient understanding and professional communication, the validation of health terms serves as a cornerstone for improving HL and DHL by facilitating patient education and ensuring that health information is clear, accurate, and coherent. This has a relevant impact on citizen engagement by leading to a better

³⁴ <https://www.snomed.org/>

understanding of diagnoses and treatments while increasing patient participation and satisfaction.

7.2 Terminology validation processes

7.2.1 Validation and verification

In the context of Terminology, the concepts of <Validation> and <Verification>³⁵ are pillars of precision and effectiveness in the quality of medical language and consequently in health communication. Both serve different purposes but in a complementary way. <Validation> should be understood as a process that ensures that the relationship between a verbal designation - monolingual or polylingual unit - and its respective concept is correctly confirmed as belonging to the area of knowledge in question and that its definition is clearly understood. Through confirmation by experts, the correctness, precision, and usability of the terminological units are ensured (Silva, 2014).

<Verification> involves confirming that terms or terminological combinations are used appropriately within a specific context of specialized written or oral communication. It focuses on ensuring the correct use of the term by verifying its proper application by established linguistic rules or within a particular domain of knowledge like medicine. This aspect of “verification” is crucial to ensure that terms are not only validated as abstract objects belonging to a conceptual system but are also used correctly in practice, thereby preventing communication failures or misunderstandings.

In summary, while validation focuses on the accuracy and appropriateness of terms in representing concepts, verification ensures that these terms are used correctly within specific contexts. Both processes are essential for effective communication in linguistics, particularly in specialized fields such as healthcare, where precise terminology is crucial for patient understanding and care. This multifaceted approach ensures that the terminological contents are not only accurate but also functional for communication within specialized domains.

7.2.2 Mediation process

The concept of <Mediation> in Terminology occurred associated with the concepts of <Validation> and <Verification> of terminological data. It highlights the methodology designed to involve the different interlocutors involved in the process, generally a group of experts. <Mediation> is primarily defined as an ethical communication process based on the responsibility and autonomy of participants, in which an independent party facilitates understanding between other parties (Guillaume-Hofnung, 2012).

This is a structured process, designed and led by the linguist/terminologist. Their role is to promote a linguistic and conceptual approach to terminological data and develop

³⁵ Concepts applied in ISO standard 9000:2015: <Validation> [3.8.13] *confirmation, through the provision of objective evidence, that the particular requirements for a specific intended use or application have been fulfilled* and <Verification> [3.8.12] *confirmation, through the provision of objective evidence, that specified requirements have been fulfilled*. These definitions served as a starting point for reflection (cf. Silva, 2014) and were adapted to terminology validation.

strategies to capture the expert's knowledge. The terminologist's task is to guide a discussion at a linguistic level, questioning the panel of experts to obtain validation of the terminological data. The terminologist relies on a validation script (Costa & Silva, 2006)³⁶, where the linguistic operations for terminological validation are described (cf. 7.3.), as well as the list of terminological data intended for discussion during the working session with the experts (Silva & Costa, 2019).

Conducting the mediation process does not involve taking sides but focuses on resolving issues related to terminology and/or concepts to organize this information more effectively. This process culminates in expert consensus, confirming that the terms are applicable due to their clarity, precision, and usability (Silva, 2014).

7.2.3 Framework for validation

The validation methodology will be carefully designed, aligning with the overarching objectives achieved through validation, the specific context of the task, and the desired outcomes. Simultaneously, it is essential to establish a structured framework that addresses a set of critical requirements. The specification of requirements guarantees the quality of the validation process, making clear objectives for obtaining adequate results designed according to users' needs.

Table 1. Requirements for terminology validation.

Specification of objectives	Purpose of validation: clarify why validation is necessary (description/standardization of terms; improving communication clarity; ensuring suitability for specific audiences). Expected outcomes: identify practical results (validated glossary/database/other resources) or enhanced communication within a particular domain.
Description of the specialized area	Definition of the domain: specify the field of expertise (medicine, law, economics, etc.). Domain organization: comprehensive and organized description, identify interdisciplinary connections with other specialized domains.
Field of application	Technical fields: terminology, lexicography, terminology management; specialized communication; translation; localization; interpretation; education; teaching; language policies; automatic translation; IT; IA.
Characterization of end users	Types of users: experts; semi-experts, researchers; students/trainees; generalists, non-experts, lay users; institutions, organizations, policy-experts. Identification of needs: ensure the adequacy of validated terminology for

³⁶ Costa, R. & Silva, R. (2006). Guião: metodologia para a investigação aplicada em Terminologia. (*Script: Methodology for research applied in terminology*) FCSH, Universidad NOVA de Lisboa (not published).

	effective application (monolingual, multilingual, cultural aspects) for communication and increasing knowledge.
Accessibility needs	Communication channels: digital platforms (databases, glossaries, etc.), app; digital formats; printed format. Types of tools: CAT tool for automatic translation; terminology extraction; AI or NLP tools; thesaurus, ontologies, etc. Inclusive design and formats: for users with special needs (assistive technologies, voice recognition, braille, etc.).
Engagement of experts	Choice of the experts: a set of professionals with in-depth knowledge of the domain and the ability to explain knowledge taking into account multidisciplinary, cultural contexts and multilingualism.

7.3 Linguistic guidelines for validation

Linguistic guidelines for terminology validation are a set of principles and procedures that help ensure that terms are linguistically and conceptually suitable for the purposes the terminology will serve and comply with previously defined requirements (cf. 7.2.3). These guidelines are especially useful in collaborative processes involving experts, terminologists, and end users.

7.3.1 Objectives for validation

In the HEREDITARY project, the validation processes will allow end users (cf. 7.3.2) to have access to a full range of terminological information about the diseases in focus, i.e. neurodegenerative diseases, in particular diseases where the relation between the gut and brain organs is at the forefront center. Regarding the terminology component, the project aims to make medical terms and concepts more accessible and understandable to the general public. This involves identifying the concepts and their respective linguistic designations, validating the terms, and reformulating or writing definitions that are understandable by non-experts in the domain. All information will be made available on the HEREDITARY platform, which will integrate a terminological database associated with other conceptual-level resources that will have the function of guaranteeing the transfer of knowledge, clarifying information, and answering user questions.

7.3.2 Stakeholders³⁷ and end users

7.3.2.1 Domain experts

a) Health professionals and their institutions: clinicians (physicians, nurses, etc.) who work in close contact with neurodegenerative and gut microbiome disorders.

³⁷ Detailed information about stakeholders in Deliverable D6.1 - Guidelines and manual for applying the Health Social Labs methodology lead by Observa.

They have mastery of scientific knowledge and terminology as well as experience in transmitting information to patients and family members. **b) Patients' association representatives:** organizations formed to support and advocate for individuals affected by the disease. Some members are **c) patient experts**, individuals who have lived experiences with specific health conditions or diseases and possess in-depth knowledge about their conditions through personal experience, self-education, or formal training.

7.3.2.2 Semi-experts

a) Health researchers and their institutions: individuals who conduct theoretical and applied research with the aim of specialization in neurodegenerative disorders and gut microbiomes. **b) Students/trainees:** individuals who are beginning to develop capabilities and learn about the domain. At different levels, researchers and students are in the process of consolidating their knowledge.

7.3.2.3 Non-experts

a) Patients: individuals receiving medical care, in this case, related to neurodegenerative or intestinal microbiome disorders. **b) Caregivers:** individuals whose job is to care for a person affected by neurodegenerative or intestinal microbiome disorder. They are those who request clear and understandable information about their health condition. **c) Lay users:** individuals who wish to be informed or seek general, simplified, and accessible information for everyday understanding. This target group of non-experts will be the privileged beneficiaries of this project.

7.3.3 Selection of terms and concepts for validation

To find the list of terminological units that must be validated by experts, the methodological approach combines semasiological and onomasiological perspectives. The first consists of compiling a textual corpus on the domain under study (cf. Section 4, Domain-specific corpus: tools and methods) and, using semi-automatic extraction tools, collecting a list of candidate terms³⁸ (cf. Section 5.5, Semi-automated terminology extraction: tools and methods). The second approach is complementary to the first, and it consists of interacting directly (cf. 7.2.2) with experts (meetings, interviews, focus groups) to collect their knowledge (Silva, 2014).

7.3.4 Orientations for terminological validation

7.3.4.1 Experts' engagement

A panel of experts is created according to criteria that vary from project to project to start the validation process. Ideally, it should include individuals whose training and experience are recognized by their peers. The terminologist supervises the experts' validation/verification tasks through the mediation process (cf. 7.2.2) which involves decision-making based on the linguistic and conceptual dimensions, to ensure the accuracy and usability of the terms. These operations include (adapt. Silva & Costa, 2019):

- a) Identifying the concept: establishing the idea or entity represented by the term;

³⁸ According to ISO standard 12616-1:2021 <candidate term> is a "string of characters that has been collected by means of term extraction but has not yet been selected as a text element to be documented in the terminological data collection".

- b) Identifying the concept's characteristics: defining the essential attributes that shape the concept;
- c) Refining verbal designations: improving term clarity and alignment with the concept;
- d) Establishing correspondence between linguistic and conceptual levels: verifying the match between a term and its concept;
- e) Identifying lexicosemantic relations: exploring relations between terms (e.g., synonymy, variation, etc.);
- f) Identifying conceptual relations: determining connections between concepts (e.g., hierarchies, associations, etc.);
- g) Reformulating or drafting definitions: creating clear, precise, and comprehensive definitions;
- h) Indicating linguistic equivalents: providing accurate translations or equivalent terms in foreign languages.

7.3.4.2 Citizen' engagement

Concepts are knowledge units that are not perceived by all in the same way, just as scientific terms also represent a high degree of difficulty in understanding for a public of semi and non-specialists and citizens in general. To involve citizens in validation, it is necessary to define collaborative strategies in which these particular users can express their opinions about terminological data. These strategies will present vulgarization and reformulation techniques about the terms and definitions, as well as simplification of medical language³⁹.

To conclude, by integrating a robust terminological validation framework through the application of validation guidelines, the HEREDITARY project can guarantee that its objectives of developing an interactive solution, aimed at different user profiles (researchers, health professionals, innovators), in which the knowledge transmitted through terminological work is capable of supporting prevention, decision-making and strengthening citizens' confidence in health matters. These objectives are achieved by making communication clearer and more precise while adopting collaborative and inclusive approaches.

³⁹ In the future deliverables D6. 5 – Citizen science and terminology: Methodology [M20] and D6. 6 – Citizen science and terminology [M48] the methodology that leads to terminological validation in the context of this project will be developed and applied aiming at scientific popularization, health literacy, and citizen science.

8 Design of a FAIR terminology resource

8.1 Introduction to FAIR principles

The FAIR principles⁴⁰ - Findability, Accessibility, Interoperability, and Reusability - have emerged as a framework to guide the management of data and ensure its broad usability in the scientific community. These principles proposed by Wilkinson et al. (2016) have become integral to research data management, fostering better data sharing, integration, and reuse. They are built upon the idea that data should be structured, documented, and stored in ways that make it easy for both humans and machines to find, access, and use it.

Findability emphasizes that data must be easy to locate. This is facilitated by persistent identifiers (such as Digital Object Identifiers or DOIs), comprehensive metadata, and a clear description of the data's content and context.

Accessibility dictates that data should be stored in a way that ensures its retrieval through well-established protocols. It also refers to data being available under clear licensing conditions and at a consistent, retrievable location.

Interoperability concerns the ability of data to be integrated with other datasets, systems, or tools. This is achieved by ensuring that data is structured in standardized formats and can be linked to external resources through common vocabularies, ontologies, or frameworks.

Reusability emphasizes that data should be reusable for future research or applications. This requires clear and rich metadata, proper documentation, and compatibility with different contexts.

In the context of terminology science, these principles are crucial as terminological data represents specialized knowledge that is used across a range of domains, languages, and systems.

8.2 Overview of the FAIR terminology paradigm

The FAIR terminology paradigm is an extension of the broader FAIR principles, applied specifically to the management of terminology resources (Vezzani, 2022). The main objective of the FAIR terminology paradigm is to ensure that terminological data is structured and maintained in ways that maximize its utility, especially in a multilingual and multidisciplinary environment.

This paradigm relies upon the complementary application of the following three ISO TC/37 SC3 standards:

⁴⁰ <https://www.go-fair.org/fair-principles/>

1. ISO 16642: 2017 *Computer applications in terminology — Terminological markup framework*⁴¹ which defines the Terminological Markup Framework (TMF) metamodel for the representation of terminological data collections.⁴²
2. ISO 12620: 2019 *Computer applications in terminology — Data categories*⁴³ (now superseded by ISO 12620: 2022 *Management of terminology resources — Data categories — Part 1: Specifications*⁴⁴ and ISO 12620: 2022 *Management of terminology resources — Data categories — Part 2: Repositories*)⁴⁵ which consistently defines the properties of data categories (such as their names and definitions) and their documentation in an open repository.
3. ISO 30042: 2019. *Management of terminology resources — TermBase eXchange (TBX)*⁴⁶, which defines the TermBase eXchange (TBX) representation format specifically designed for the exchange of multilingual terminological data.

As discussed in Vezzani et al. (2023: 241), these three standards inherently intersect with the four FAIR principles, particularly regarding the interoperability and reusability of terminological resources.

ISO 16642: 2017 emphasizes a modular methodology for creating interoperable terminological data collections. Its primary aim is to “[...] facilitate cooperation and to prevent duplicate work, [...] as well as for sharing and exchanging data” (ISO 16642:2017, vi). This focus positions the standard as strongly aligned with the principle of ‘interoperability.’ The terminological data collections outlined in ISO 16642:2017 are designed to accommodate diverse data categories, which must be identified and organized across various environments.

ISO 12620: 2019 details a framework for “creating, documenting, harmonizing and maintaining data category specifications in a data category repository” (ISO 12620:2019, 1). It outlines how data categories should be effectively identified and accessed, making it particularly relevant to the FAIR principles of ‘findability’ and ‘accessibility.’

Meanwhile, ISO 30042:2019 serves as a model for representing structured terminological data in XML. This standard is designed to “[...] support various types of processes involving terminological data, including analysis, descriptive representation, dissemination, and exchange in various computer environments” (ISO 30042:2019, vi). Its key objective is to enable the exchange of terminological data for diverse purposes, closely aligning it with the ‘reusability’ principle of the FAIR framework.

Together, these standards provide the foundation for implementing the FAIR terminology paradigm. This concept was initially developed during the creation of TriMED, a multilingual and multipurpose medical terminology resource (Vezzani et al., 2018;

⁴¹ <https://www.iso.org/standard/56063.html>

⁴² At present, ISO 16642: 2017 is being updated by ISO TC 37/SC 3/WG 3: <https://www.iso.org/standard/87351.html>

⁴³ <https://www.iso.org/standard/69550.html>

⁴⁴ <https://www.iso.org/standard/79078.html>

⁴⁵ <https://www.iso.org/standard/79018.html>

⁴⁶ <https://www.iso.org/standard/62510.html>

Vezzani and Di Nunzio, 2020a, 2020b). It was subsequently fully integrated into the development of two additional multilingual resources: CAMEO (CommerciAl terMinology rEsource), targeting international trade terminology (Vezzani and Di Nunzio, 2022), and DITTO (Disarmament International Treaty TerminOlogy), focusing on international disarmament (Vezzani et al., 2022).

8.3 Current state of research

The adoption of the abovementioned ISO TC/37 standards has significantly influenced research in terminology management and semantic interoperability, shaping tools and methodologies for creating and exchanging terminological resources. In this section, we present a review of the most recent papers using these standards.⁴⁷

For example, Vacalopoulou et al. (2019) demonstrated how standardized data categories improve system interoperability in Greek museums, emphasizing the need for domain-specific adjustments. Warburton and Wright (2020) highlighted ISO 12620's flexibility, which allows the customization of data categories for diverse terminological needs. Benoît (2020) explored how linguistic registers in ISO 12620 enhance the accessibility of medical resources, illustrating the added value of standards in domain-specific platforms like healthcare and cultural heritage.

Innovative applications of these standards extend to knowledge organization systems. Arndt and Runnwerth (2021) proposed concept-oriented practices, integrating TBX to establish controlled vocabularies, while Mihăescu (2021) introduced conceptual hierarchies in termbases aligned with TMF principles. Corporate implementations also underscore the standards' importance. Fišer and Witt (2022) leveraged TBX in the Termportalen portal within the CLARIN framework, and Meisinger et al. (2022) refined ISO 12620 for terminological databases. Warburton et al. (2021) stressed the standards' utility in corporate terminology management and their relevance to emerging professional roles like technical communicators.

Efforts to bridge TBX with Linked Data environments further demonstrate the evolving role of these standards. TBX, primarily designed for data exchange, has been adapted for the Semantic Web using RDF. Cimiano et al. (2015) introduced the TBX2RDF framework with OntoLex-Lemon to publish terminologies as Linked Data. Subsequent work by Speranza et al. (2020) and Piccini et al. (2023) enriched TBX resources with semantic annotations. Bellandi et al. (2024) automated TBX-to-RDF conversion, while Martín-Chozas and Declerck (2022) extended OntoLex-Lemon for unaddressed terminological data. Reineke and Romary (2019) proposed mapping TBX and SKOS to improve interoperability, and Nunzio and Vezzani (2021) advocated for abstract data modeling to ensure terminological resources adapt seamlessly to different formats while preserving their conceptual and linguistic dimensions.

⁴⁷ An extended and detailed version of the literature review presented here is presented in Vezzani et al. (2023: 236-240).

These studies collectively highlight the potential of the abovementioned ISO standards in advancing terminology management research.

8.3.1 Data entry interface

To create a new concept entry, users must first select a specialized domain by accessing the subject field dropdown menu. This menu includes a list of domains sourced from EuroVoc, the multilingual and multidisciplinary thesaurus developed by the European Union. Once a domain is chosen, a new concept entry can be initiated by clicking the Add Concept Entry button, which automatically generates a unique identification number in the concept field.

Additional information at the concept level can be entered by expanding the corresponding section using the Show/Hide icon. At this stage, users may:

1. Specify a more detailed subdomain.
2. Establish relationships with other concept entries by inputting their identification numbers into the appropriate relational fields (e.g., subordinate, superordinate, comprehensive, or partitive).




Figure 2. FAIRterm 2.0, concept level (screenshot).

From the concept level, users can also add language-specific sections by selecting a language from the Select Language to Add dropdown. The list of available languages adheres to the ISO 639 standard for language codes (ISO 639:2023). For each language section, users can populate the following data categories: (1) definition, (2) external Cross-Reference, (3) source, (4) notes.

Figure 3. FAIRterm 2.0, language level (screenshot).

Figure 3. FAIRterm 2.0, language level (screenshot).

Additionally, term sections can be appended to each language section by selecting the Add Term Section option. Among all sections, the term section offers the most comprehensive set of data categories for detailed input, including: (1) designation, (2) usage, (3) part of speech, (4) gender, (5) number, (6) type, (7) context, (8) external cross reference, (9) source, (10) register, (11) collocation, and (12) notes.

The interface design accommodates an unlimited number of language and term sections, arranged vertically. To manage screen space effectively, sections can be expanded or collapsed using the Show/Hide feature. Each section operates independently, with its own Update button to save entered information without affecting other sections.

8.3.2 Data consultation interface

The Data Consultation interface is similar in structure to the Data Entry interface but offers a more concise and streamlined layout. It lacks editing capabilities and is primarily designed for viewing compiled data. Both interfaces include a Search Terms bar, enabling users to locate specific terms quickly. Search functionality can include or exclude definitions as required.

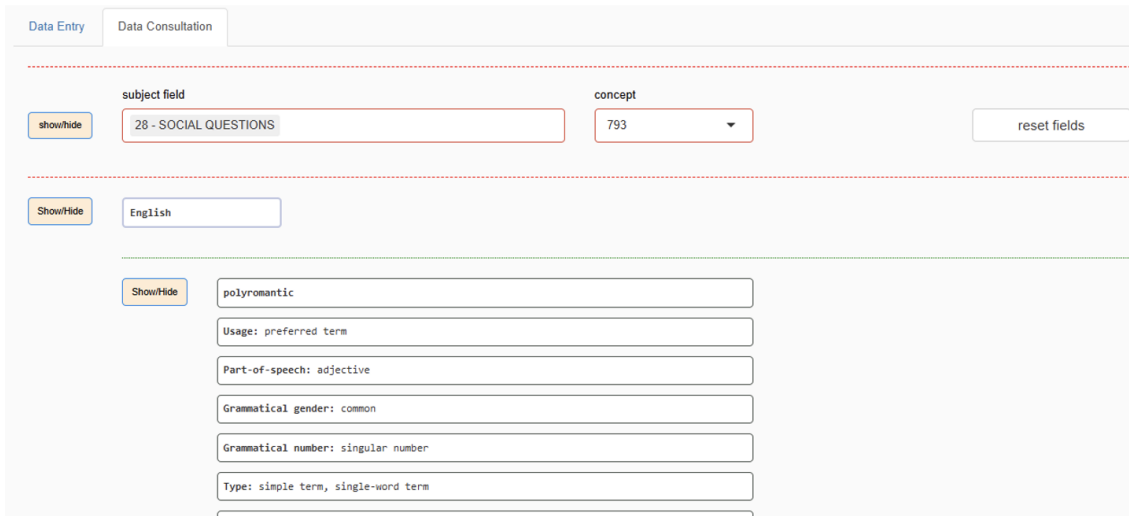


Figure 4. FAIRterm 2.0, data consultation interface (screenshot).

Finally, FAIRterm 2.0 allows users to reuse the compiled terminological data by downloading it in two formats: a tabular format (tsv) for simplified processing and analysis, and a TBX format that adheres to the ISO 30042:2019 standard, ensuring interoperability and compliance with international terminological standards. This dual export capability enhances the system's flexibility and facilitates integration into various workflows.

9 Conclusions

This deliverable presents a comprehensive approach to establish a unified methodology for managing terminological data, ensuring that terminology is accessible and comprehensible to all stakeholders involved in organizational, structuring, and communication processes. This work is built upon recognized medical resources, forming the foundation for constructing a coherent and consistent corpus.

Our methodology emphasizes terminology quality, which is why it is grounded in standardization tools and formats. The proposed steps are designed to ensure quality across corpus building, terminology extraction, and validation processes, all of which are essential for effective communication. The FAIRterm design is intended to support high-quality data, as will be demonstrated in the upcoming stages of our work.

By applying established theoretical frameworks for conceptual and lexical relations, while also addressing the dual dimensions of terminology, we establish a solid foundation for a robust methodology.

10 Annexes

Number	Name
Annex 1	<Amyotrophic Lateral Sclerosis (disorder)> in SNOMED CT
Annex 2	Rule-based corpus queries for lexical markers identification

REFERENCES

Key	Reference
American Medical Association, 2021	American Medical Association. (2021). Guidelines for Patient Communication.
Arndt, et al., 2021	Arndt, Susanne, and Mila Runnwerth. (2021). "Linked Vocabularies for Mobility and Transport Research." In Metadata and Semantic Research: 14th International Conference, MTSR 2020, Madrid, Spain, December 2–4, 2020, Revised Selected Papers 14, edited by Emmanouel Garoufallou and María Antonia Ovalle-Perandones, 168–179. Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-71903-6_17
Baker, et al. 2006	Baker, P., Hardie, A., & McEnery, T. (2006). A Glossary of Corpus Linguistics (In the series Glossaries in Linguistics ed.). Edinburgh University Press. https://doi.org/10.1515/9780748626908
Barrière, 2006	Barrière, Caroline. (2004). Building a concept hierarchy from corpus analysis. In Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, Vol.10, Issue 2 (pp. 241-263). https://doi.org/10.1075/term.10.2.05bar
Bellandi, et al., 2024	Bellandi, Andrea, Di Nunzio, Giorgio Maria, Piccini, Silvia, & Vezzani, Federica. (2024). LemonizeTBX: Design and Implementation of a New Converter from TBX to OntoLex-Lemon. DHQ: Digital Humanities Quarterly, 18(2).
Benoît, 2020	Benoît, Gerald. (2020). "Searching Covid-19 by Linguistic Register: Parallels and Warrant for a New Retrieval Model". In Proceedings of the Association for Information Science and Technology 57 (1): 1-11. https://doi.org/10.1002/pa2.246
Bonato, et al., 2024	Bonato, V., Di Nunzio, G. M., & Vezzani, F. (2024). A Novel Approach to Semic Analysis: Extraction of Atoms of Meaning to Study Polysemy and Polyreferentiality. Languages, 9(4), 121. https://doi.org/10.3390/languages9040121
Bourigault, et al., 2001	Bourigault, Didier, Christian Jacquemin, and Marie-Claude L'Homme. (2001). Recent Advances in Computational Terminology. John Benjamins. https://www.jbe-platform.com/content/books/9789027298164
Carvalho, 2018	Carvalho, S. (2018). A terminological approach to knowledge organization within the scope of endometriosis: the EndoTerm project. PhD Thesis. Universidade Nova de

Key	Reference
	Lisboa/Communauté Université Grenoble Alpes, https://run.unl.pt/handle/10362/49745
Carvalho, et al., 2018	Carvalho, S., Costa, R. & Roche, C. (2018). The Role of Conceptual Relations in the Drafting of Natural Language Definitions: an Example from the Biomedical Domain, in I. Kernerman, S. Krek, (eds.). Proceedings of the LREC 2018 Workshop Globalex 2018 – Lexicography & WordNets. Miyazaki: European Language Resources Association (ELRA), 10-16. ISBN 979-10-95546-28-3.
Carvalho, et al., 2023	Carvalho, S., Wermuth, C. & Costa, R. (2023). Definitions in SNOMED CT through the lens of Terminology: from formal to textual. In: Di Nunzio, G., Costa, R. & Vezzani, F. (eds.), Proceedings of the 2nd International Conference on Multilingual digital terminology today. Design, representation formats and management systems (MDTT 2023), Lisbon, Portugal, June 29-30, 2023, https://ceur-ws.org/Vol-3427/paper1.pdf
Chute & Cohn, 2019	Chute, C. G., & Cohn, S. P. (2019). Health Informatics: Practical Guide for Healthcare and Information Technology Professionals. Springer.
Chute, 1998	Chute, C. G. (1998). The Copernican era of healthcare terminology: a re-centering of health information systems. Proceedings. AMIA Symposium (pp. 68–73).
Cimiano, et al., 2015	Cimiano, Philipp, John P. McCrae, Víctor Rodríguez-Doncel, Tatiana Gornostay, Asunción Gómez-Pérez, Benjamin Siemoneit, and Andis Lagzdins. (2015). “Linked Terminologies: Applying Linked Data Principles to Terminological Resources.” In Proceedings of the ELex 2015 Conference (pp. 504–517).
Cimino, 1998	Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. Methods of information in medicine, 37, 4-5, (pp. 394–403).
Codagnone, et al., 2019	Codagnone, M. G., Spichak, S., O'Mahony, S. M., O'Leary, O. F., Clarke, G., Stanton, C., Dinan, T. G., & Cryan, J. F. (2019). Programming Bugs: Microbiota and the Developmental Origins of Brain Health and Disease. Biological psychiatry, 85(2), 150–163. https://doi.org/10.1016/j.biopsych.2018.06.014
Condamines & Rebeyrolle, 2001	Condamines, A., & Rebeyrolle, J. (2001). Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB): Method and Results. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), Recent Advances in Computational

Key	Reference
	Terminology (Vol. 2, pp. 127-148). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Costa, 2006	Costa, Rute. (2006). Plurality of Theoretical Approaches to Terminology. In P. Heribert (Ed.), Linguistic Insights. Studies in Language and Communication (Vol. 36). Berlin - Bern: Peter Lang Verlag.
Costa, 2013	Costa, Rute. (2013). Terminology and specialized lexicography: two complementary domains, <i>Lexicographica</i> , 29, (pp. 29-42).
Costa, 2017	Costa, Rute. (2017). Les normes en terminologie. Que faire des synonymes ?. Normes linguistiques et terminologiques : conflits d'usages [Eds. Danielle Candel et Hélène Ledouble] Cahiers de Lexicologie, 2017-1, n° 110, Paris: Classiques Garnier (pp. 45-57). ISBN 978-2-406-07056-6.
Cruse & Anthony, 2000	Cruse, A. & David Anthony (2000). <i>Meaning in Language: An Introduction to Semantics and Pragmatics</i> . Oxford University Press.
Cruse, 2002	Cruse, A. (2002). Hyponymy and Its Varieties. In R. Green (Ed.), <i>The semantics of relationships</i> (pp. 3 - 21). Springer Science+Business Media Dordrecht.
Di Nunzio, et al., 2021	Di Nunzio, Giorgio Maria, and Federica Vezzani. (2021). One Size Fits All: A Conceptual Data Model for Any Approach to Terminology. (arXiv:2112.06562). arXiv. https://doi.org/10.48550/arXiv.2112.06562
Di Nunzio, et al., 2023	Di Nunzio, Giorgio Maria, Stefano Marchesin, and Gianmaria Silvello. (2023). A Systematic Review of Automatic Term Extraction: What Happened in 2022? . <i>Digital Scholarship in the Humanities</i> 38, no. Supplement_1 (pp. i41–47). https://doi.org/10.1093/llc/fgad030
Drouin, et al., 2015	Drouin, Patrick, Natalia Grabar, Thierry Hamon, and Kyo Kageura. (2015). Introduction to the Special Issue: Terminology across Languages and Domains. In <i>Terminology</i> 21, no. 2 (pp. 139–50). https://doi.org/10.1075/term.21.2.01dro
Drouin, et al., 2018	Drouin, Patrick, Natalia Grabar, Thierry Hamon, Kyo Kageura, and Koichi Takeuchi. (2018). Computational Terminology and Filtering of Terminological Information: Introduction to the Special Issue. In <i>Terminology</i> 24, no. 1 (pp. 1–6). https://doi.org/10.1075/term.00010.dro
Duclos, et al., 2014	Duclos, C., Burgun, A., Lamy, J.-B., Landais, P., Rodrigues, J.-M., Soualmia, L. and Zweigenbaum, P. (2014). Medical Vocabulary, Terminological Resources and Information Coding in the Health Domain, in Venot, A., Burgun, A. and

Key	Reference
	Quantin, C. (eds.) Medical Informatics, e-Health. Health Informatics. Springer (pp. 11-41).
Fellbaum, 1998	Fellbaum, Christiane. (1998). WordNet: An Electronic Lexical Database. MIT Press.
Fišer & Witt, 2022	Fišer, Darja, and Andreas Witt. (2022). CLARIN: The Infrastructure for Language Resources. Berlin: De Gruyter.
Frieden, 2017	Frieden, T. R. (2017). Evidence for Health Decision Making — Beyond Randomized, Controlled Trials. The New England Journal of Medicine.
Ginsburg & Phillips, 2018	Ginsburg, G. S., & Phillips, K. A. (2018). Precision medicine: From science to value. Health Affairs.
Grosjean, et al., 2011	Grosjean, J., Merabti, T., Dahamna, B., Kergourlay, I., Thirion B., Soualmia, L. F. and Darmoni, S. J. (2011). Health Multi-Terminology Portal: a semantics added-value for patient safety. Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety, Studies in Health Technology and Informatics, Volume 166, (pp. 129-138).
Guillaume-Hofnung, 2012	Guillaume-Hofnung, M. (2012). La médiation, 6e édition, Paris: PUF Coll. Que sais-je?, n° 2930.
Halskov & Barrière, 2010	Halskov, J., & Barrière, C. (2010). Web-based extraction of semantic relation instances for terminology work. In A. Auger, & C. Barrière (Eds.), Probing Semantic Relations: Exploration and identification in specialized texts (Vol. 23, pp. 20-42). Amsterdam Philadelphia: John Benjamins B.V.
Hanh, et al., 2023	Tran, Hanh Thi Hong, Matej Martinc, Jaya Caporusso, Antoine Doucet, and Senja Pollak. (2023). The Recent Advances in Automatic Term Extraction: A Survey. arXiv https://doi.org/10.48550/arXiv.2301.06767
Hersh, 2018	Hersh, W. (2018). Information Retrieval: A Health and Biomedical Perspective. Springer.
Institute for Healthcare Improvement, 2022	Institute for Healthcare Improvement. (2022). Teach-Back Method.
ISO 1087:2019	International Organization for Standardization. (2019). Terminology work — Vocabulary — Part 1: Theory and application. (ISO Standard No. 1087:2019). https://www.iso.org/standard/62330.html

Key	Reference
ISO 12616-1:2021	International Organization for Standardization. (2021). Terminology work in support of multilingual communication — Part 1: Fundamentals of translation-oriented terminography. (ISO Standard no. 12616-1:2021). https://www.iso.org/standard/72308.html
ISO 12620:2019	International Organization for Standardization. (2019). Management of terminology resources – Data category specifications. (ISO Standard no. 12620:2019). https://www.iso.org/standard/69550.html
ISO 12620-1:2022	International Organization for Standardization. (2022). Management of terminology resources Management of terminology resources — Data categories — Part 1: Specifications. (ISO Standard no. 12620-1:2022). https://www.iso.org/standard/79078.html
ISO 12620-2:2022	International Organization for Standardization. (2022). Management of terminology resources Management of terminology resources — Data categories — Part 2: Repositories. (ISO Standard no. 12620-2:2022). https://www.iso.org/standard/79018.html
ISO 9000:2015	International Organization for Standardization. (2015). Quality management systems — Fundamentals and vocabulary. (ISO Standard no. ISO 9000:2015). https://www.iso.org/standard/45481.html
ISO 16642:2017	International Organization for Standardization. (2017). Computer applications in terminology – Terminological markup framework. (ISO Standard no. 16642:2017). https://www.iso.org/standard/56063.html
ISO 30042:2019	International Organization for Standardization. (2019). Management of terminology resources— Term-Base eXchange (TBX). (ISO Standard no. 30042:2019). https://www.iso.org/standard/62510.html
ISO 704:2022	International Organization for Standardization. (2022). Terminology work — Principles and methods. (ISO Standard no. 704:2022). https://www.iso.org/standard/79077.html
ISO/FDIS 5078	International Organization for Standardization. (2024). Management of terminology resources — Terminology extraction. (ISO/FDIS Standard no. 5078). https://www.iso.org/standard/81917.html
ISO/TR 14639-1:2012	International Organization for Standardization. (2012). Health informatics — Capacity-based eHealth architecture roadmap. Part 1: Overview of national eHealth initiatives.

Key	Reference
	(ISO/TR no. 14639-1:2012). https://www.iso.org/standard/54902.html
Martín-Chozas & Declerck, 2022	Martín-Chozas, Patricia, and Thierry Declerck. (2022). “Representing Multilingual Terminologies with OntoLex-Lemon”. In Proceedings of the 1st International Conference on Multilingual Digital Terminology Today, vol. 3161, edited by Giorgio Maria Di Nunzio, Geneviève Marie Henrot, Maria Teresa Musacchio and Federica Vezzani. CEUR-WS. https://ceur-ws.org/Vol-3161/#short1
Martins, et al., 2024	Martins, A., Velez, L., Isabel, L. N., Giordano, A. P., Semedo, H., Vital, C., Silva, R., Coelho, P., & Londral, A. (2024). A conversational agent for enhanced self-management after cardiothoracic surgery. <i>International Journal of Medical Informatics</i> , 192, 1-8. ISSN (pp. 1872-8243).
Meisinger, et al., 2022	Meisinger, Nino, Thorsten Trippel, and Claus Zinn. (2022). “Increasing CMDI’s Semantic Interoperability with schema.org”. Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2714–2720. https://aclanthology.org/2022.lrec-1.290
Meyer & Mackintosh, 2000	Meyer, I. and Mackintosh, K. (2000). When terms move into our everyday lives: An overview of de-terminologization. <i>Terminology</i> , vol. 6:1. Amsterdam: John Benjamins, (pp. 111-138).
Meyer, 2001	Meyer, I. (2001). Extracting Knowledge-Rich contexts for terminography: a conceptual and methodological framework. In D. Bourigault, C. Jacquemin, & M.-C. L’Homme (Eds.), <i>Recent Advances in Computational Terminology [Natural Language Processing, 2 ed.]</i> , (pp. 279 - 302). Amsterdam / Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/nlp.2.15mey
Mihăescu, 2021	Mihăescu, Manuela. (2021). “Ressources et traitements automatiques des informations terminologiques”. <i>Revue Internationale d’études en Langues Modernes Appliquées</i> 14 (pp. 15–30).
Mikolov, et al., 2013	Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean. (2013). Efficient Estimation of Word Representations in Vector Space. In 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings , edited by Yoshua Bengio and Yann LeCun. http://arxiv.org/abs/1301.3781
Norman & Skinner, 2006	Norman, C. D., & Skinner, H. A. (2006). eHealth Literacy: Essential Skills for Consumer Health in a Networked World. <i>Journal of Medical Internet Research</i> , 8(2), e9.

Key	Reference
Nuopponen, 1994	Nuopponen, Anita. (1994). Concept Systems for Terminological Analysis. Doctoral Dissertation. University of Vaasa, Finland.
Nuopponen, 2005	Nuopponen, Anita (2005). Concept Relations. An Update of a Concept Relation Classification. In B. N. Madsen & H. Erdman Thomsen (Eds.), Terminology and Content Development: TKE 2005, 7th International Conference on Terminology and Knowledge Engineering (pp. 127-138). Litera.
Nuopponen, 2014	Nuopponen, Anita (2014). Tangled Web of Concept Relations. Concept relations for ISO 1087-1 and ISO 704. Terminology and Knowledge Engineering 2014, Jun 2014, Berlin, Germany. 10 p. hal-01005882
Nuopponen, 2018	Nuopponen, Anita (2018). Terminological Concept Systems. In Languages for Special Purposes: An International Handbook, edited by John Humbley, Gerhard Budin and Christer Laurén, Berlin, Boston: De Gruyter Mouton (pp. 453-468). https://doi.org/10.1515/9783110228014-023
Nuopponen, 2022	Nuopponen, Anita (2022). Conceptual relations: From the General Theory of Terminology to knowledge bases. In P. Faber & M.-C. L'Homme (Eds.), Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge (pp. 63-86). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.23.03nuo
Nutbeam, 2008	Nutbeam, D. (2008). The evolving concept of health literacy. Social Science & Medicine.
Piccini, et al., 2003	Piccini, Silvia, Federica Vezzani, and Andrea Bellandi. (2023). TBX and 'Lemon': What perspectives in terminology? In Digital Scholarship in the Humanities 38. Supplement_1 (pp. i61-i72).
Ramos & Costa, 2024	Ramos, M., & Costa, R. (2024). Applying Text Mining Methods to Construct a Domain Ontology from Definitions. Umanistica Digitale (UD) – ISSN 2532-8816 [Forthcoming]
Ramos, 2020	Ramos, M. (2020). Knowledge Organization and Terminology: application to Cork. PhD Thesis. Chambéry : Université Savoie Mont Blanc (NNT :) https://hal.science/tel-03106436 ; Lisboa : Universidade NOVA de Lisboa http://hdl.handle.net/10362/111722
Ramos, et al., 2019	Ramos, M., Costa, R., Roche, C. (2019). Dealing with specialized co-text in text mining: Verbal terminological collocations. In TOTh 2019 Terminologie & Ontologie : Théories et Applications. Terminologica. Le Bourget du Lac : Presses Universitaires Savoie Mont Blanc. (hal-02891157)

Key	Reference
Rector, 1999	Rector, A. L. (1999). Clinical terminology: why is it so hard? <i>Methods of information in medicine</i> , 38(4-5), 239–252.
Rector, et al., 2011	Rector, A. L., Brandt, S., & Schneider, T. (2011). Getting the terminology right: Terminology maintenance and evolution. <i>Journal of the American Medical Informatics Association</i> .
Reineke, et al., 2019	Reineke, Detlef, and Laurent Romary. (2019). “Bridging the gap between SKOS and TBX”. <i>Edition - Die Fachzeitschrift Für Terminologie</i> 19 (2): 19-27. https://hal.inria.fr/hal-02398820
Santos & Costa, 2015	Santos, C. & Costa, R. (2015). Domain specificity: Semasiological and Onomasiological knowledge representation. In H. J. Kockaert & F. Steurs (Eds.), <i>Handbook of Terminology</i> (Vol. 1, pp. 153-179). Amsterdam: John Benjamins Publishing Company.
Silva & Costa, 2019	Silva, R., Costa, R. (2019). Accéder aux connaissances des experts par l’entremise de la médiation en Terminologie. In M. D. Gioia, & M. Marcon (Eds.), <i>L’essentiel de la médiation : Le regard des sciences humaines et sociales</i> (pp. 105-121). Ed. Peter Lang.
Silva, 2014	Silva, R. (2014). <i>Gestão de Terminologia pela Qualidade - Processos de validação</i> . Ph.D. thesis. Universidade Nova de Lisboa – Faculdade de Ciências Sociais e Humanas, Lisbon.
Silva, et al., 2023	Silva, R., von Hafe, F., Azevedo, S., & Londral, A. (2023). Popularizing Terminology Using Social Networks: Keeping Citizens Informed About Value in Health Care. <i>CEUR Workshop Proceedings</i> (CEUR-WS.org), ISSN 1613-0073.
Silverman, et al., 2013	Silverman, J., Kurtz, S., & Draper, J. (2013). <i>Skills for Communicating with Patients</i> . CRC Press.
Sørensen, et al. 2012	Sørensen, K., et al. (2012). Health literacy and public health: A systematic review. <i>BMC Public Health</i> .
Sparck Jones, 1972	Sparck Jones, Karen. (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. In <i>Journal of Documentation</i> 28, no. 1 (pp. 11–21). https://doi.org/10.1108/eb026526
Speranza, et al., 2020	Speranza, Giulia, Maria Pia Di Buono, Johanna Monti, and Federico Sangati. (2020). “From Linguistic Resources to Ontology-Aware Terminologies: Minding the Representation Gap”. In <i>Proceedings of the Twelfth Language Resources and Evaluation Conference</i> (pp. 2503–2510). https://aclanthology.org/2020.lrec-1.305

Key	Reference
Topol, 2019	Topol, E. (2019). <i>Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again</i> . Basic Books.
Vacalopoulou, et al., 2019	Vacalopoulou, Anna, Stella Markantonatou, Katerina Toraki, and Panagiotis Minos. (2019). "Open-Access Resource for the Management and Promotion of Greek Museums with Folk Exhibits". In <i>Strategic Innovative Marketing and Tourism</i> , edited by Androniki Kavoura, Efsthios Kefallonitis and Apostolos Giovanis (pp. 129–137). Switzerland : Springer International Publishing. https://doi.org/10.1007/978-3-030-12453-3_15
Vezzani & Di Nunzio, 2020a	Vezzani, Federica, and Giorgio Maria Di Nunzio. (2020a). Methodology for the Standardization of Terminological Resources: Design of TriMED Database to Support Multi-register Medical Communication. In <i>Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication</i> 26 (2) (pp. 265-297).
Vezzani & Di Nunzio, 2020b	Vezzani, Federica, and Giorgio Maria Di Nunzio. (2020b). On the Formal Standardization of Terminology Resources: The Case Study of TriMED. In <i>Proceedings of the twelfth International Conference on Language Resources and Evaluation (LREC 2020)</i> . European Language Resources Association (ELRA) (pp. 4903–4910).
Vezzani & Di Nunzio, 2022	Vezzani, Federica, and Giorgio Maria Di Nunzio. (2022). Elaborazione e gestione di (meta) dati terminologici. In <i>Risorse e strumenti per l'elaborazione e la diffusione della terminologia</i> , edited by Elena Chiocchetti and Natascia Ralli (pp. 152-168). Eurac Research.
Vezzani, 2021	Vezzani, Federica. (2021). La ressource FAIRterm : entre pratique pédagogique et professionnalisation en traduction spécialisée. In <i>Synergies Italie</i> 17 (pp. 51-64).
Vezzani, 2022	Vezzani, Federica. (2022). <i>Terminologie numérique : conception, représentation et gestion</i> . Bern: Peter Lang.
Vezzani, et al., 2018	Vezzani, Federica, Giorgio Maria Di Nunzio, and Genevieve Henrot. (2018). TriMED: A Multilingual Terminological Database. In <i>Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)</i> . European Language Resources Association (ELRA) (pp. 4367-4371).
Vezzani, et al., 2022	Vezzani, Federica, Giorgio Maria Di Nunzio, and Sara Silecchia. (2022). La fraseologia dei trattati internazionali di disarmo: La risorsa terminologica DITTO. <i>Umanistica Digitale</i> 14 (pp. 91-117). https://doi.org/10.6092/issn.2532-8816/14796

Key	Reference
Vezzani, et al., 2023	Vezzani, Federica, Giorgio Maria Di Nunzio, and Rute Costa. (2023). ISO standards for terminology resources management: Are they FAIR enough? In <i>Digital Translation 1 (2)</i> (pp. 233-252).
Warburton & Wright, 2020	Warburton, Kara, and Sue Ellen Wright. (2020). A Data Category Repository for Language Resources. In <i>Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences</i> , edited by Antonio Pareja-Lora, María Blume, Barbara C. Lust and Christian Chiarcos (pp. 69-98). Cambridge, Massachusetts: MIT Press.
Warburton, 2021	Warburton, Kara. (2021). <i>The Corporate Terminologist</i> . Amsterdam/Philadelphia: John Benjamins.
Wermuth & Verplaetse, 2018	Wermuth, M. C. and Verplaetse, H. (2018). Medical terminology in the Western world: current situation, in Alsulaiman, A. and Allaithy, A. (eds.) <i>Handbook of Terminology: Terminology in the Arab World</i> . Amsterdam: John Benjamins (pp. 83-108).
Whetzel, et al., 2011	Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T. and Musen, M.A. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. <i>Nucleic Acids Res.</i> 2011 Jul;39 (Web Server issue): W541-5. Epub 2011 Jun 14.
Wilkinson, et al., 2016	Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. In <i>Scientific Data 3 (1)</i> (pp. 1-9).
Winston, et al., 1987	Winston, M., Chaffin, R., & Hermann, D. (1987). <i>A Taxonomy of Part-Whole Relationships</i> . Cognitive Science. https://www.researchgate.net/publication/245104866_A_Taxonomy_of_Part-Whole_Relationships
WHO, 2013	World Health Organization. (2013). <i>Health Literacy: The Solid Facts</i> . Geneva: WHO
WHO, 2016	World Health Organization. (2016). <i>Health literacy and patient-centered communication</i> . Geneva: WHO
WHO, 2019	World Health Organization. (2019). <i>International Classification of Diseases, 11th Revision (ICD-11)</i> . Geneva: WHO.
Zauner, 1902	Zauner, Adolfe. (1902). <i>Die romanischen Namen der Körperteile: Eine onomasiologische Studie</i> , K.B. Hof- und Universitäts-Buchdruckerei von F. Junge, Erlangen.