Evaluation Task Annotation Protocol

Version 1 - 2026/01/22

Federica Vezzani[1], Giorgio Maria Di Nunzio[2], Vanessa Bonato[1], Gianmaria Silvello[2]

[1] Department of Linguistic and Literary Studies, University of Padua (Italy)
[2] Department of Information Engineering, University of Padua (Italy)

Document Structure

Overview

The DEfinition and Term Extraction CHallenge (DETECH) is a new evaluation challenge that focuses on the automatic extraction of medical terms from domain-specific specialized documents and the generation of natural language definitions for medical concepts. By integrating these tasks, DETECH seeks to promote research on data-driven medical terminology, while establishing a framework to evaluate automatic methods for identifying and defining specialized concepts.

The challenge is organized within the HEREDITARY (HetERogeneous sEmantic Data integratIon for the guT-brAin inteRplaY) project,[1] which is a European-founded project aimed at advancing knowledge on the gut-brain interplay and its relation with a wide range of neurodegenerative diseases, including Parkinson's disease, Alzheimer's disease, and Multiple Sclerosis.

From a terminological perspective, the acquisition of domain-specific knowledge involves identifying terms and concepts, as well as defining concepts used in specialized communication. The challenge is divided into two subtasks:

- **Task A - Term Extraction**: Participants are provided with English abstracts retrieved from PubMed,[2] and are asked to identify both single-word terms and multi-word terms considered relevant to the gut-brain interplay field of research.
- **Task B - Definition Generation**: Participants are asked to produce English natural language definitions for the concepts designated by the extracted terms. Definitions can be produced by relying on corpus-based evidence, or generated using automatic text generation methods.

In the following sections, we describe the annotation protocol adopted to create the DETECH datasets for Task A (Term Extraction) and Task B (Definition Generation). We detail the criteria used to identify domain-relevant terms in the documents, the procedure for mapping terms to concepts, and the principles followed to formulate intensional definitions, including specific conventions and special cases encountered during data preparation.

[1] https://hereditary-project.eu
[2] https://pubmed.ncbi.nlm.nih.gov

Task A – Term Extraction

Participants will receive a collection of PubMed documents concerning the gut-brain interplay. Each document consists of the title and the abstract of a scientific article. These documents were retrieved by performing two distinct queries, using the keywords: 1) “gut microbiota” AND “mental health”, and 2) “gut microbiota” AND “Parkinson”.[3]

The objective is to identify domain-relevant terms from documents (titles and abstracts), by following a text-based approach. Terms can be either single-word terms or multi-word terms. In this task, terms correspond to lexical units relevant to the gut-brain interplay field of research. These lexical units may or may not be documented in existing biomedical terminology resources and ontologies listed in the “Consulted Resources” section below. In particular:

- We consider “gut microbiota”, “synucleinopathy”, “Akkermansia Muciniphila”, “dysbiosis” and “neurodegenerative disorder” to be domain-relevant terms. In terminology resources and ontologies, these terms respectively appear with the exact same sequence of characters.
- We also consider “whole food microbiota modulating intervention”, “FEP patient”, “microbe-derived compound”, “probiotic food research” and “halobiont environment” to be domain-relevant terms. In terminology resources and ontologies, these lexical units would not appear with the exact same character sequences. However, they convey domain-relevant information within the analyzed documents.

[3] The query was performed on PubMed in May 2025.

Task B – Definition Generation

The objective of this subtask is to produce natural language definitions for the concepts designated by the extracted terms, based on corpus-based evidence or using automatic text generation techniques. Participants are asked to produce terminological definitions in the form of intensional definitions, following the guidelines provided in ISO 1087 (2019) and ISO 704 (2022).

Identification of Concepts

In this subtask, a concept is considered a "unit of knowledge created by a unique combination of characteristics" (ISO 1087: 2019). For example:

- We consider <Gut microbiota>,[4] <Synucleinopathy>, <Akkermansia muciniphila>, <Dysbiosis>, and <Neurodegenerative disorders> to be concepts.
- We consider "whole food microbiota modulating intervention", "FEP patient", "microbe-derived compound" to be domain-relevant lexical units in Task A. However, at the conceptual level, we usually consider them as composed concepts, that could be split into different concepts as follows:
    - "whole food microbiota modulating intervention" could be split into <Whole food>, <Microbiota>, and <Intervention>;
    - "FEP patient" could be split into <First-episode psychosis> and <Patient>;
    - "microbe-derived compound" could be split into <Microbe> and <Compound>.

  For the lexical units which refer to composed concepts, there are two options:

  1) defining serapately the different concepts (i.e., <First-episode psychosis> and <Patient>)
  2) or defining the concept as a whole <FEP patient>.

  In the training set, you will find examples of both cases. For the evaluation, both solutions will be accepted.

[4] According to the annotation convention used in the community, we distinguish concepts from terms as follows: concepts are capitalized and written between single chevrons (<>), whereas terms are in lowercase and between double quotation marks ("").

Formulation of Intensional Definitions

An intensional definition is a “definition that conveys the intension of a concept by stating the immediate generic concept and the delimiting characteristics” (ISO 1087: 2019). The generic concept is the “concept in a generic relation that has the narrower intension”. A delimiting characteristic is defined as an “essential characteristic used for distinguishing a concept from related concepts”. The following example shows the intensional definition of the concept <Microbiome>:

> *biome that consists of the microorganisms, their genomes (genes), and the surrounding environment where the microorganisms reside*

This intensional definition is formulated according to the following rules:

- The immediate generic concept, <Biome>, is the first element mentioned in the definition, followed by the set of delimiting characteristics of the concept.
- The defined concept <Microbiome> is linguistically designated by a term that, from a grammatical standpoint, is a noun. Accordingly, the immediate generic concept in the definition is also a noun.
- The definition follows the typographic conventions recommended in ISO 704 (2022), as it is “a statement in the form of an incomplete sentence without a full stop” and it is not preceded by an article.

The intensional definition, moreover, “shall not contain characteristics that belong logically to its generic concepts or specific concepts”. Based on this consideration, we provide a correct example (A) and an incorrect example (B) of the intensional definition of the concept <Gut microbiome>:

A. microbiome that resides in the gut
B. microbiome that resides in the gut, that consists of the microorganisms, their genomes (genes), and the surrounding environment where the microorganisms reside

Particular Cases

In the biomedical domain under investigation, the definitions of certain concepts are not universally agreed upon. An example of this concerns the concept <Fecal microbiota transplantation>, whose definition may vary depending on the considered regulation

(Merrick et al., 2020). In particular, this affects the selection of the immediate generic concept to be included in the intensional definition.

To formulate the intensional definition of the concept <Fecal microbiota transplantation>, a text-based approach was adopted, considering <Transpantation> as the immediate generic concept:

transplantation of fecal microbiota from a healthy individual to a recipient

Consulted Resources

For the formulation of the intensional definitions included in the training set, multiple resources were consulted. Priority was given to specialized biomedical resources and ontologies that provide information relevant for acquiring specialized biomedical knowledge, such as definitions of concepts and a classification or hierarchical organization of concepts. Additionally, scientific articles from PubMED were consulted as a source of specialized knowledge, being a reference database for biomedical literature. The consulted resources are listed below, in order of priority:

1. NCI Thesaurus (NCIT).[5] Managed by the National Cancer Institute (NCI) Enterprise Vocabulary Services (EVS), it covers biomedical terminology, providing information on synonyms, definitions of concepts and a hierarchical organization of concepts.
2. Medical Subject Headings (MeSH Terms).[6] MeSH is the controlled vocabulary of the National Library of Medicine. It provides information on biomedical concepts, such as the definitions of concepts and a hierarchical organization of concepts.
3. Chemical Entities of Biological Interest (ChEBI).[7] It is an ontology specifically focused on molecular entities and chemicals. It includes both definitions of concepts and their hierarchical organization within the resource.
4. Unified Medical Language System (UMLS).[8] UMLS integrates information concerning biomedical terminology contained in different biomedical resources, providing for instance definitions from resources such as NCI, MeSH and SNOMED CT.
5. Diagnostic and statistical manual of mental disorders Fifth Edition, Text revision - DSM-5-TR (American Psychiatric Association, 2022). It provides a classification of

[5] http://purl.obolibrary.org/obo/ncit.owl
[6] https://www.ncbi.nlm.nih.gov/mesh/
[7] http://purl.obolibrary.org/obo/chebi.owl
[8] https://uts.nlm.nih.gov/uts/umls/home

mental disorders, accompanied by diagnostic criteria for each classified disorder. It includes relevant information such as diagnostic features and differential diagnoses.

6. International Classification of Diseases 11th Revision (ICD-11).[9] It is an international system maintained by the World Health Organization. It provides a classification of medical conditions, accompanied by descriptions of the concepts.
7. Gene Ontology (GO).[10] It provides information on genes of organisms, genes products and biological processes. It includes definitions of concepts and their hierarchical organization.
8. Domain-relevant scientific articles retrieved from PubMed.

[9] https://icd.who.int/en/
[10] http://purl.obolibrary.org/obo/go.owl

References

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.). American Psychiatric Association Publishing.

ISO 1087. (2019). *Terminology work and terminology science - Vocabulary*. International Organization for Standardization. https://www.iso.org/standard/62330.html

ISO 704. (2022). *Terminology work - Principles and methods*. International Organization for Standardization. https://www.iso.org/standard/79077.html

Merrick, B., Allen, L., Masirah M Zain, N., Forbes, B., Shawcross, D. L., & Goldenberg, S. D. (2020). Regulation, risk and safety of Faecal Microbiota Transplant. *Infection Prevention in Practice*, *2*(3), 100069. https://doi.org/10.1016/j.infpip.2020.100069