Methodology for the standardization of terminological resources: design of TriMED database to support multi-register medical communication

Federica Vezzani[1] and Giorgio Maria Di Nunzio[2,3]

[1] *Dept. of Linguistic and Literary Studies, University of Padua, Italy*

federica.vezzani@phd.unipd.it

[2] *Dept. of Information Engineering, University of Padua, Italy*

[3] *Dept. of Mathematics, University of Padua, Italy*

giorgiomaria.dinunzio@unipd.it

Abstract

Terminology standardization reflects two different aspects involving the meaning of terms and the structure of terminological resources. In this paper, we focus on the structural aspect of standardization and we present the work of re-modeling TriMED, a multilingual terminological database conceived to support multi-register medical communication. In particular, we provide a general methodology to make the termbase compliant to three of the most recent ISO/TC 37 standards. We focus on the definition of i) the structural meta-model of the resource, ii) the provided data categories and its Data Category Repository, and iii) the TBX format for its implementation.

Keywords: Term base design; Term management; FAIR Terminology; Terminology Standardization; Medical Terminology

1 Introduction

Terminology standardization has long been acknowledged as a fundamental requirement when working with special languages (Wüster 1971; Strehlow 1988; Sager 1998). This process reflects two different aspects related to terminology: semantic and structural.

From a semantic perspective, the standardization of terminology entails selecting a designation expressing a given concept (or meaning). To ensure efficient communication in specialized contexts, there is a tendency towards linguistic uniformization. Therefore, the notion of standardization points to an idea of "harmonization" of the meaning of terms and related concepts (Cabré Castellví 1997). The main tasks are to 1) create a shared common consensus on the use of technical terms within a community of experts, and 2) disambiguate their meaning to encourage monosemy: a term should have a single univocal meaning (Magris et al. 2002). Unambiguous communication is encouraged whenever concepts have the same meaning for all the participants in an interaction process, at the national or international level. However, communication can be difficult at the international level because concepts may not coincide and the conceptual systems differ from language to language (Felber 1980). The standardization of the meaning of terms becomes particularly important in contexts where the transmission of information is "vital". Communication in the healthcare domain, for example, is often characterized by opaque lexicon that has to be disambiguated (Rouleau 2003; Castro et al. 2007). As stated in Galinski and Nedobity 1988, the standardization of terminology is not always necessary, sensible, or desirable in all fields of human endeavors. However, standardization becomes indispensable when the safety of human beings depends on a specialized activity, as in the case of medical emergency services.

Concerning the structural perspective, the standardization of terminology, considered in this context as a collection of terminological data, reflects the standards established to help organize and manage structured data for linguistic resources. In Stührenberg 2012, the author distinguishes between *de jure* standards, that is guidelines defined by organizations that have been assigned to this task, and *de facto* standards which are non-endorsed specifications by a standards body as they have achieved greater popularity compared to similar specifications. Herein, we refer to *de jure* standards and, in particular, to the terminology guidelines published by the International Organization for Standardization (ISO),[1] Technical Committee ISO/TC 37 working on *Language and Terminology*. The latter provides a large number of standards for the "management of terminology resources" (sub-committee 3) and "language resource management" (sub-committee 4). In addition, standardization efforts are increasingly adopted in international projects such as CLARIN (Common Language Resources and Technology Infrastructure),[2] aiming to harmonize formats and standards for language resources and technology. The main goal of this terminology standardization process is to provide structural models in order to dissociate the content from the structure of terminological data. Structured and standardized terminological data can therefore be reused in different terminology management systems.

Two key concepts guide the management of standard terminologies: interoperabil-

1. `https://www.iso.org/home.html`
2. `https://www.clarin.eu`

ity and reusability. Interoperability emerges from the structural level of terminological resources, so that "interoperable" resources share the same structural model. Instead, reusability appears at the terminological data level which, if standardized, can be used in different terminological resources in conformity to such standard (see section 2 for a literature review).

Such aspects are significant not only for proper implementation of a terminological resource, but also in a broader sense of 'Open Science', where both key concepts of "interoperability" and "reusability" are two fundamental keystones of the FAIR (Findability, Accessibility, Interoperability and Reusability) principles (Wilkinson et al. 2016) within the European Open Science Cloud (ESOC).[3] These guidelines underline the need for "data curation" as a good practice to optimize the organization and sharing of research data (Palmer et al. 2013; Corti et al. 2019).

1.1 Our contribution

Within the context of a FAIR terminology management, our work focuses on the structural standardization of an existing multilingual database named TriMED (Vezzani, Di Nunzio, and Henrot 2018). This database, conceived specifically to support multi-register medical communication, addresses some of the problems that may arise from the complexity of medical terminology, such as the transmission and comprehension of medical information. Due to the dualism between semantic and structural standardization, this paper only treats the second structural aspect. It describes a general methodology to implement a FAIR terminological resource, so that stored terminological data are capable of responding to such principles. To our knowledge, the TriMED resource is the first implementation of the FAIR principles to terminological data. The resource is designed to comply with

3. `https://www.eosc-portal.eu`

three current ISO standards for the implementation and management of terminological resources. In particular, we present the:

1. standardization of the structure of TriMED based on the Terminological Markup Framework (TMF) meta-model (ISO-16642 2017a);

2. description of its Data Categories Specifications through the implementation of the TriMED Data Category Repository (ISO-12620 2019);

3. implementation of the resource in the TermBase eXchange (TBX) format (ISO-30042 2019);

4. Web application to access the data stored in the Termbase.

The remainder of this paper is organized as follows: in Section 2, we present the initiatives of both semantic (2.1) and structural standardization (2.2) assuming medical terminology as our working domain. In this wide panorama, we will also provide a description of the terminological variation management initiatives (2.3). In Section 3, we describe the application of the above-mentioned current standards to the TriMED resource by focusing on the 1) structural meta-model 3.1, 2) data categories 3.2, and 3) TBX implementation 3.3. In Section 4, we present our Web Application and the description of the data collection visible to users accessing the TriMED Web application (4.4). Finally, in Section 5, we present our conclusions and describe our future work.

2 Related Works

As stated in the introduction, the standardization of terminology reflects both on the content of the information and on the structure of the resources. We therefore make a distinction between semantic and structural standardization. By "semantic standardization" we mean all the initiatives carried out to standardize the designation expressing concepts that convey technical meaning and to create a common shared consensus on the use of terms. Instead, by "structural standardization" we mean all the activities aiming to standardize the structure of terminological resources. The literature review proposed in the following sub-sections focuses mostly on our field of work - the medical domain. Nevertheless, it is worth noting that the initiatives and general methodologies for the structural standardization of terminological resources described in Section 2.2 are domain-independent.

Finally, this overview aims to provide a comprehensive panorama of the available terminological resources for the medical domain. Although partially or indirectly related to the structural standardization issue, we present the initiatives related to the management of terminological variation as another important aspect when dealing with special languages and the implementation of terminological resources.

2.1 Semantic Standardization Initiatives

Semantic standardization of scientific terminology, as a language for special purposes (LSP), has a long-standing tradition that dates back to Linné's system of scientific nomenclature (Galinski and Nedobity 1988). Behind this effort stems the need to guarantee univocity and unambiguity of technical terms in order to optimize communication, in particular, among experts working in their specialized sub-domains. To this purpose, nomenclatures, vocabularies, terminologies and codes have been developed to enable effective communication among medical experts and to facilitate recording patient data (De Quirós, Otero, and Luna 2018; Wermuth and Verplaetse 2019).

In the field of anatomy, for example, the process of semantic standardization goes back to 1895, when the first Latin anatomical nomenclature was published as Basiliensia Nomina Anatomica. Seven revisions ensued (Jenaiensia Nomina Anatomica 1935, Parisiensia Nomina Anatomica 1955, Nomina Anatomica 2nd to 6th edition 1960–1989). The last revision, Terminologia Anatomica, (TA) compiled by the Federative Committee on Anatomical Terminology and approved by the International Federation of Associations of Anatomists, was published in 1998 (Kachlik et al. 2008). The work contains international reference standard nomenclature in the field of human anatomy, collecting terminological indications on approximately 7,500 macroscopic human anatomical structures. Pathology is another medical branch involving the study and diagnosis of diseases which has been subject to standardization. In this context, the International Classification of Diseases (ICD)[4] is the international standard diagnostic tool used in epidemiology, health management and for clinical purposes. Such tool is managed by the World Health Organization (WHO) which is the directing and coordinating authority for health within the United Nations System. The latest ICD-11[5] version was released on 18 June 2018 to allow Member States to prepare for its implementation, including the translation of ICD into their national languages. ICD-11 was submitted to the 144th Executive Board Meeting in January 2019, to the Seventy-second World Health Assembly in May 2019. Once endorsed, Member States will start reporting using ICD-11 on 1 January 2022. Another resource to standardize detailed clinical information is the Systematized Nomenclature of Human and Veterinary Medicine - Clinical Terms (SNOMED-CT),[6] considered as the most comprehensive, multilingual clinical healthcare terminology in the world (Benson 2012). This tool serves the following purposes: i) to enable more accurate documentation of patient care, ii) to enhance the reporting and analysis of medical outcomes and iii) to facilitate

4. `https://www.who.int/classifications/icd/en/`
5. `https://icd.who.int/en/`
6. `http://www.snomed.org`

semantic interoperability by capturing clinical data in a standardized, unambiguous and granular manner (Bowman 2005; Lee et al. 2013). Moreover, the Medical Subject Headings (MeSH) Thesaurus[7] is configured as the most widely used resource for the indexing and retrieval of the biomedical literature task. When firstly introduced in 1960, Frank B. Rogers, the director of the National Library of Medicine, acknowledged this to be "the adoption of standard topical subheadings for cataloging books, as well as for indexing periodical articles" (Lipscomb 2000). This controlled vocabulary contains standardized terms used to search for topics in articles indexed in MEDLINE[8] and NLM'S PubMed.[9] Finally, in 1986, the National Library of Medicine launched a long-term research and development project to build the Unified Medical Language System[10] (UMLS) (Lindberg, Humphreys, and McCray 1993). The compendium gathers a large number of health and biomedical vocabularies and standards (Association et al. 2015) conceived, in particular, to enhance semantic interoperability and linguistic uniformity (Ingenerf, Reiner, and Seik 2001).

To conclude, whether it is thesaurus, terminologies, nomenclatures or classification systems, such medical resources have been implemented and are now being used in order to avoid semantic ambiguities in the medical context and to encourage a proper inter- and intra-linguistic transmission of medical information.

2.2 Structural Standardization Initiatives

The structural standardization of domain-specific terminological resources is conceived as a collection of *de jure* standards developed by the International Organization for Standardization (ISO),[11] concerning specifications on their correct design and implementation.

7. https://meshb.nlm.nih.gov/search
8. https://www.medline.com
9. https://www.ncbi.nlm.nih.gov/pubmed/
10. http://bit.ly/2IWDfuQ
11. https://www.iso.org/home.html

This notion is strictly related to the FAIR principles (Wilkinson et al. 2016) of the European Open Science Cloud (ESOC)[12] designed to promote the "FAIRness" (Findability, Accessibility, Interoperability and Reusability) of research data. Published in 2016, these guidelines refer to scholarly output, including data, metadata and workflows surrounding them. The aim is to address the lack of widely-shared, clearly stated, and broadly applicable best-practices on the publication of data generated by scientific research (Da Silva Santos et al. 2016).

For the medical domain, in particular, the international standard (ISO-171171 2018), developed by the Technical Committee 215, working on *Health informatics*, defines universal and specialized characteristics of health terminological resources. It only refers to tools designed primarily for clinical concept representation. The structural standardization of these medical resources espouses the onomasiological approach of the General Theory of Terminology that dates back to Wüster 1979: the resources should have a conceptual orientation, thus be configured as ontologies, that is collections of clinical concepts and corresponding medical terms which designate their meaning.

This kind of structural representation is also shared and spread by ISO-16642 2017a developed by the ISO TC/37 sub-committee 3 working on "Management of Terminology Resources". The standard defines the Terminological Markup Framework (TMF) meta-model: the basic idea is that a terminological resource collects a set of terminological entries, each referring exclusively to a single concept. Concepts can be represented in multiple languages through terms designating the meaning of the concepts themselves. The main goal of such abstract structural model is to encourage interoperability, so that "interoperable" resources should share the same structural model. The current version of the standard dates back to 2017, and is a second edition following the previous one

12. `https://www.eosc-portal.eu`

published in 2003 (for the historical development of the standard see Romary 2001b; Romary et al. 2006).

In the management of a FAIR terminology, even the concept of reusability of terminological data becomes important in the structural standardization process. Reusability implies that the terminological data should have a standard format to be interchanged (imported or exported) among different terminological resources that support the format. Several authors highlight the need for exchange of terminological data, and describe more or less obsolete formats, such as: the Electronic Terminology Interchange Format (ETIF) (Melby 1995), the Machine Readable Terminology Interchange Format (MARTIF) (Melby, Schmitz, and Wright 2001), and the TermBase Exchange (TBX) format (Schmitz 2012; Melby 2015).[13]

Except for the ISO-171171 2018, the application of such standards to favor interoperability and reusability is domain-independent. With reference to the scientific domain, two important initiatives of structural standardization are worth mentioning: TermSciences and Meriterm. The TermSciences[14] initiative deals with the construction of a multi-purpose and multi-lingual terminological database using a variety of source vocabularies compiled by major French research institutions and which were initially intended to be used for indexing and cataloguing scientific literature (Khayari et al. 2006). This was the first public initiative to implement the Terminological Markup Framework meta-model for the interchange of computerized lexical data. The authors refer to the previous version of the ISO 16642 on Computer Applications in Terminology published in 2003, which proved suited for concept modeling since it allowed to organize the original resources by

13. Regarding the TBX format, the authors refer to the first 2008 edition. In the following section (in particular sec 3.3) we refer to the latest 2019 version.

14. `http://www.termsciences.fr`

concepts and to associate the different terms for a given concept. The same onomasiological approach has been used to develop the Medical Enduser and Reference Interface Terminology[15] (MERITERM), a project by Roumier et al. 2011. The group proposes a Medical Interface Terminology linked to large existing linguistic resources and to a broad array of international medical classifications according to terminology standards (the TMF meta-model) and Linked Data Principles (Warnier et al. 2012; Cardillo et al. 2013).

2.3 Terminological Variation Management Initiatives

Despite the numerous initiatives for semantic standardization of terminology in the medical domain, it is important to note that such efforts inevitably collide with the natural phenomenon of terminological variation. In fact, many studies focus on different aspects and types of term variation that frequently occur in medical language such as, among others: "reversible" variants (Bowker and Hawkins 2006); synonyms, inflectional and derivational variants (Llanos et al. 2016); lexical variants (Delavigne 2017). One of the most important phenomena in terminological variation that occurs in the medical domain is register variation, or diastratic variation (Coseriu 1969). Based on the typology proposal by Freixa 2006, this phenomenon is attributable to functional causes when the level of language specialization among interlocutors has to be adapted (see for example Picton and Dury 2017).

In this context, numerous resources have been specifically implemented to manage the terminological variation in the medical domain and, in particular, to bridge the register differences occurring in the doctor-patient dialogue (Jucks and Bromme 2007; Deuster et al. 2008; Castro et al. 2007). These resources are conceived for non-experts to facilitate the communication and the transmission of information in the healthcare domain. The

15. https://www.meriterm.org

aim is to improve understanding of medical terms rather than standardize their meaning. For example, the Consumer Health Vocabulary Initiative (Zeng and Tse 2006) has led to the development of a useful resource [16] for the English language that allows the translation of technical terms into a consumer-friendly language. In parallel, the initiative of Cardillo, Serafini, and Tamilin 2009 has led to the development of the Italian Consumer-oriented Medical Vocabulary,[17] that shows the different ways patients and healthcare consumers in general express and think about health topics. In this context, the Spanish Ministry of Economy and Competitiveness funded another project which has also led to the development of the terminological database VariMed.[18] This resource focuses on terminological variation in the medical field, providing pragmatic information to facilitate communication between healthcare professionals and patients (Sánchez and Velasco 2013). Another project for patient consultation is CureHunter Discovery Engine,[19] a fully unified and integrated numeric index of all known drugs, biologically active agents, diseases and empirical statements on all effective clinical outcomes published in the United States National Library of Medicine.[20] The database is designed to focus on definition, variation and on treatments - diseases relation. It also provides easy-to-read documents for each disease.

From a multilingual viewpoint, several initiatives have been carried out at the European level. The European Commission (DG III) published the Multilingual Glossary of Popular and Technical Medical Terms.[21] The project was implemented by the Department of Applied Linguistics of the Heymans Institute of Pharmacology and Mercator School in 1995-2000. The resource groups together nine glossaries, in nine official languages of the

16. http://bit.ly/2L3tSw4
17. http://bit.ly/2L31Fpp
18. http://varimed.ugr.es
19. http://www.curehunter.com
20. https://www.nlm.nih.gov
21. http://bit.ly/2UEMxxK

European Union, of 1830 scientific and popular medical terms about medicinal product package inserts. In 2018, the Terminology Coordination Unit (TermCoord) of the European Parliament in collaboration with the Paris-Diderot University, the University of Granada and the University of Padua, engaged in the YourTerm MED[22] project consisting in the realization of a structured linguistic database based on the cognitive structure of the Frame-Based Terminology (Faber 2015). This multilingual tool is designed to facilitate communication between healthcare professionals on mission and their patients. In particular, it strives to meet the actual terminological needs of Médecins sans Frontières/Doctors Without Borders (MSF) during their international missions. Physicians can therefore consult a multilingual resource on site and take care of patients.

Without claim to completeness, the above-mentioned initiatives do not aim at the semantic or structural standardization of terminology, but rather at completing the panorama of currently available resources designed to manage terminological variation within the medical domain.

3 TriMED Standardization

In this wide landscape collecting medical resources carried out by medical authorities for terminological standardization and linguistic resources for terminological variation, we introduce the work of remodeling an existing multilingual termbase for the medical domain named TriMED (Vezzani, Di Nunzio, and Henrot 2018). In particular, we focus on the structural standardization of the resource to fulfill the international needs expressed by the FAIR principles (described in Section 2.2) of the European Open Science Cloud (ESOC).[23] The rationale behind this work of terminological data curation is to provide

22. http://bit.ly/2VqClOO

23. https://www.eosc-portal.eu

findable, accessible, interoperable and reusable data (Wilkinson et al. 2016) to provide a multi-purpose and accurately implemented terminological infrastructure for the medical domain. To our knowledge, this is the first application that explicitly addresses the problem of the FAIR principles to terminological data. As stated in section 2.2, TermSciences and MERITERM projects are the only two initiatives that have implemented rigorous application of international standards. However, these projects refer to the previous version of the TMF structural meta-model published in 2003. In this context, the innovative contribution introduced herein through the notion of "FAIR terminology" consists in the application of all of the latest terminology management standards to guarantee, unlike other resources that are not structured according to international guidelines, an optimal and efficient (re)use of the terminological data provided.

Before describing our structural standardization pipeline, we introduce the TriMED resource, conceived to tackle different problems related to the complexity of medical language, such as diastratic variation and specialized translation, as mentioned in section 2.3. In particular, the project's main goal is to support multi-register communication in the medical domain. Its originality lies in taking into account different user profiles in order to meet different information needs through a single access point. TriMED collects structured terminological records and provides a wide range of linguistic information based upon the user identification in three languages: Italian, English and French. The resource was designed in order to:

1. Help patients (or, in general, non-experts) to understand opaque technical terms providing their equivalent expressed in a more familiar register;

2. Support scientific translators both in the "decoding" (understanding) and "transcoding" (translating) health information (Jammal 1999) from a source language into a

target language;

3. Create a unique access point for healthcare professionals by integrating standard medical terminologies, ontologies and classification systems.

A terminological record is a tool to record, in a structured set of fields, the terminological data for a specialized concept (Pavel and Nolet 2001). A terminological resource is commonly organized around a collection of terminological records gathering relevant information about technical terms. The choice of which data related to a term are to be shown to the user depends on the purpose of the resource. As stated in Pitar et al. 2011, to represent the specificity of a particular area of knowledge, the model of terminological record should be adapted to each work domain. In general, current initiatives in database design highlight a strong propensity to record multiple information on terms. For example, the ARTES database[24] is a multi-domain multilingual resource, whose model of record provides a wide variety of information in order to support the writing and translation tasks when working with languages for specific purposes (Pecman 2018). Another example of database that collects multiple information on terms is provided by the EcoLexicon resource,[25] devoted specifically to the environment domain (León-Araúz, Reimerink, and Faber 2019).

In this context, our project is in line with general tendencies to meet the multi-register and multilingual purposes. The model of terminological record provided in TriMED considers a wide range of data categories to respond to three different users needs (see Section 3.2 for the list of the TriMED data categories, and Section 4 for the presentation of the user interaction with the Web application). The terminological records in TriMED are structured internally in the application as *tibbles* ('modern' tables in the R programming

24. https://artes.app.univ-paris-diderot.fr/artes-symfony/web/app.php/fr
25. http://ecolexicon.ugr.es

language (Müller and Wickham 2019)) where each row contains the information about a term and each column a linguistic feature of that term. Therefore, one of the objectives of this work is to design a general methodology, described in the following sections, to re-model a plain table-based terminological dataset into a new data model that complies with the three ISO standards. This structural standardization has the advantage to guarantee: 1) interoperability between resources providing a structural meta-model in conformity with other Terminological Markup Languages (TML) compliant to the ISO-16642 2017b and 2) reusability of the TriMED terminological data structured according the standard format Term Base Exchange (TBX) (ISO-30042 2019) which can be exchanged (exported or imported) in several terminology management systems and computer assisted translation tools.[26]

3.1 Structural Meta-Model

To represent the TriMED database structure, we have referred to the ISO standard 16642 about *Terminological Markup Framework* (TMF) meta-model (ISO-16642 2017b). TMF is an international standard that provides a framework for the representation of Terminological Data Collections (TDCs), such as multilingual terminological databases, in eXstensible Markup Language (XML). The TMF standard is based on two levels of abstraction. The first and most abstract level concerns a description of the meta-model that supports analysis, design and exchanges at a very general level. The meta-model is therefore independent from any specific implementation or software. At the second abstract level, the data model considers data categories that may be associated within meta-model levels and which are specific to each TDCs (see sub-section 3.2). This standard therefore does not represent a particular format, but rather a meta-model based on the traditional concept-oriented view

26. https://www.tbxinfo.net/tbx-support/

of a terminological entry dating back to Wüster's early works (Picht and Schmitz 2001): a concept is described in n languages and is designated by m terms for each language.

In particular, the structure of the meta-model is configured as follows:

- A Terminological Data Collection (TDC) contains any number of Terminological Entries (TEs).
- Each TE refers to a single concept that can therefore be represented in n languages in the Language Sections (LSs).
- For each language, there are m Term Sections (TSs) containing the terms that, in that particular language, describe the concept and all the related features.
- Each TS can contain any number of Term Component Section (TCS) providing information about parts of a term such as morphemes, phonemes, syllables, or single words from a multiword term.

As shown in Figure 1, this standard proposes adopting the hub and spoke model (Van Campenhoudt 2017) and distinguishes the hierarchical levels (Romary 2001a) to which different data categories can be associated: 1) "conceptual" data common to all languages; 2) language-specific data; 3) term-specific data.

In TriMED, a single concept, which is exclusive for a terminological entry and defined by its identifier, can be expressed by n languages. Language sections have their own identifier and include the definition expressing the related concept, as well as the language code defined by the ISO standard 639.[27] Thereafter, for each language, all the terms (TSs) expressing such concept are associated with their related terminological data (not shown in Figure 1 for space reasons). Finally, in case of a multi-word term, the properties of parts of the term are grouped in the TCS. Given the traditional concept-oriented approach

27. https://www.iso.org/iso-639-language-codes.html

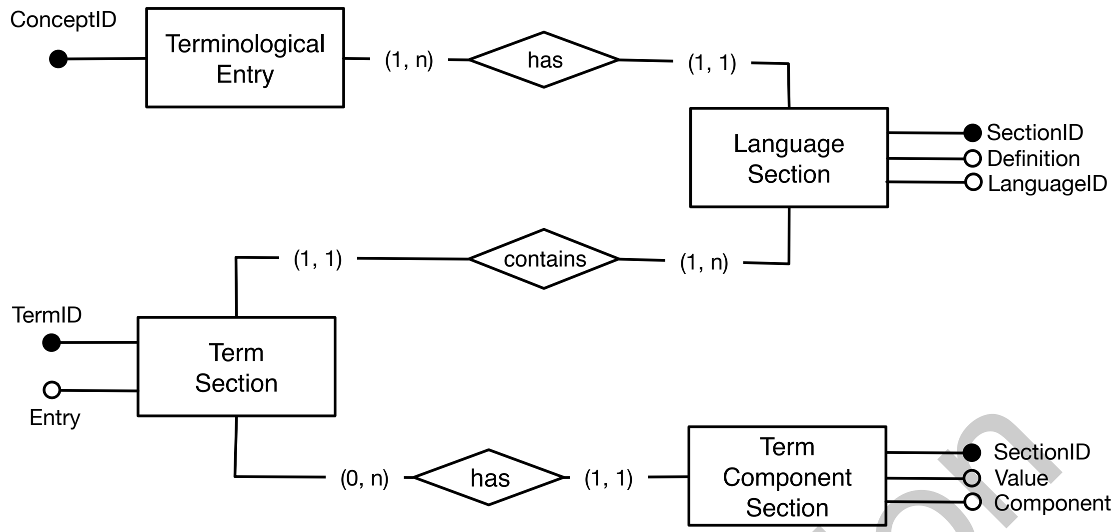

Figure 1: The Entity Relationship (ER) schema of the TMF meta-model

of the proposed terminological meta-model, one could claim how it is possible to manage the natural phenomenon of terminological variation in this kind of representation. In the following sections, we highlight how the structural model proposed can also combine the needs of terminological representation that come from more recent register and corpus-based approaches to terminology (Temmerman 2000; Cabré Castellví 2003; L'Homme 2004).

3.2 Data Categories

The four main nodes of the meta-model can be accompanied by a large number of data categories depending on the purpose of the resource. A data category is a class of information that forms part of a data collection or annotation scheme for a given language resource (for example, /definition/, /part of speech/[28] etc.) (ISO-12620 2019).

As described at the beginning of this section, the TriMED resource provides a wide range of information to satisfy the information needs of i) patients, ii) translators, and iii)

28. For compliance with the standard the data categories are presented between slashes.

healthcare professionals. In Table 1 we illustrate a list of the data categories provided in the record model for the three working languages. Additional administrative information, specifically related to the record, is also given, in particular: transaction type (origination, modification), responsibility, date, user suggestion and notes. The choice of data categories and information to be shown to the user were guided by the desire to provide a resource that can 1) help patients understand medical information correctly, considering the aspect of diastratic terminological variation, 2) support the translator in the specialized translation process providing a framework on the syntactic, semantic and phraseological behavior of the source and target term, and 3) provide healthcare professionals with a single access point to conusult other terminologies, nomenclatures or international classification codes generally used by experts (Section 4 illustrates a detailed description of which data categories are visible to each user).

To frame all the data categories associated to a term and the concept it designates, we refer to the ISO standard 12620 on *Data Category Specifications* (ISO-12620 2019). This international standard classifies all the data categories that can be associated to the TMF meta-model levels that include: 1) the "conceptual" data common to all languages; 2) the language-specific data; 3) the term-specific data. For their description and implementation, the standard refers to a Data Category Repository named DatCatInfo.[29] This resource replaces the previous ISOcat Repository conceived, in particular, to describe language resource metadata and support interoperability and reusability of terminological data (Kemps-Snijders et al. 2008; Broeder, Schuurman, and Windhouwer 2014; Windhouwer and Schuurman 2014). The new DatCatInfo repository collects a list of all the data categories available for terminology and data category specifications, being a complete and formal representation of the data categories (for example, name, definition,

29. http://www.datcatinfo.net/

Table 1: TriMED Data Categories

Field	Terminological Data
Morphology	part of speech, grammatical gender, grammatical number, derivative forms
Phonetics	IPA transcription
Etymology	derivation, composition
Variation	orthographic variant, abbreviation, full form, acronym
Semantics	definition, semic analysis, synonym, hyponym, hypernym
Phraseology	phraseological unit, collocation
Pragmatics	context of use
Register	common name, scientific name, international classification codes
Domain	subject field, subdomain
References	source (URLs)

examples, comments, etc.).

3.2.1 TriMED Data Category Repository

The novelty of this standard consists in the possibility for all implementers to create their own specific Data Category Repository (DCR) for a particular resource. In this way, defining a clear framework to specify, manage, and use data categories improves the interoperability of terminology resources. To this end, we have designed the TriMED Data Category Repository which gathers all the data category specifications, that is the description of all the terminological data we provide in our model of terminological record. The Web application was implemented using the Shiny R package (Chang 2015) and it is available online.[30] To our knowledge, this is the first DCR implemented according to

30. `http://purl.org/trimed/dcr`

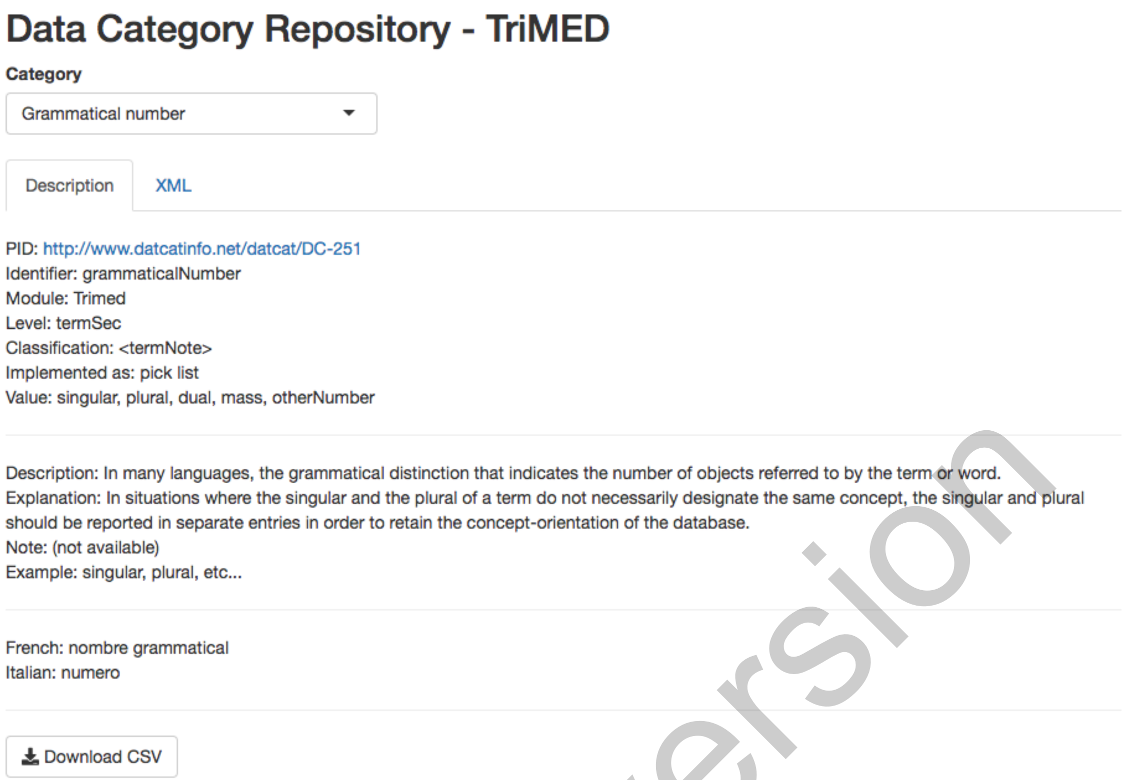

Figure 2: TriMED Data Category Repository

this standard, and it collects 42 data category specifications. As shown in Figure 2, the information provided for each data category are:

- A unique and persistent identifier (PID), that is a Uniform Resource Identifier (URI) that provides direct web access to the data category specification in the online repository.
- A unique and stable mnemonic identifier of the data category which should not include spaces in-between words, because it is used in encoding environments as an element or attribute value. In the example provided, this identifier is ‘grammaticalNumber’.
- The TBX Data Category Module to which the category refers (see Section 3.3).

- The level of the TMF meta-model (concept, language and term) to which the present data category is associated in our resource. The example illustrates the category /grammatical number/ associated to the term section.

- The content typology of the data category, that is, the types of information that the data category allows for its implementation, such as a pick list or a text string.

- A set of enumerated values that the data category can have if implemented as a pick list. The example shows admissible values for the /grammatical number/ which are single, plural, dual, mass or other number.

- A clear definition of the concept of the data category. For /grammatical number/ data category, this information corresponds to the following description: *the grammatical distinction that indicates the number of objects referred to by the term or word.*

- Further explanations and notes about data category.

- Some examples of use of data category, such as *singular* or *plural.*

- The translation of the canonical name of the data category into the other working languages of the resource. In the example provided *nombre grammatical* for French and *numero* for Italian.

It is worth noting that, to ensure traceability, in our DCR the information provided for the PID of the data category refers to the corresponding URI on DatCatInfo, for categories that have already been documented in the repository, as suggested by the standard. For example, the PID of the data category /grammatical number/ is the same provided in DatCatInfo, that is http://www.datcatinfo.net/datcat/DC-251. There are also a restricted number of data categories which are exclusive to our resource and therefore not illustrated in DatCatInfo. These terminological data categories include: 1) semic analysis, 2) hypernym, 3) hyponym, 4) subdomain, 5) ICPC2 code, 6) ICD10 code, 7) SNOMED CT term, 8) MeSh term and 9) conceptual sphere. For these categories, we provide a PID that corresponds to the URI of the related page in our TriMED DCR. For example, in the case of semic analysis, that is the decomposition of the meaning of technical terms (lexematic or morphological unity) into minimal units of meaning (Rastier 2009), the PID provided is: http://purl.org/trimed/dcr/dc/dc-1

The creation of our own data category repository enables us to disambiguate the meaning of some categories of data. By consulting the resource, users can find all the necessary information to understand the provided data. The TriMED DCR embodies guidelines to correctly compile the terminological records in terms of data "consistency". In particular, by consistency we mean rigorous naming of data categories represented within the resource (Wright 2001): the set of values of each data category is defined by the DCR itself. The category /grammatical number/ has therefore a single identifier, that is the machine readable string of characters 'grammaticalNumber' for which the system does not accept any other variant. Finally, as suggested by the standard, the user can export the information provided in the DCR both in CSV and XML formats.

3.3 TBX Implementation

After defining the structural meta-model of the terminological data collection and our data categories, we describe in this section the implementation format chosen for the TriMED resource. For its representation, we have referred to the ISO standard 30042 dealing with *TermBase eXchange* (TBX) format (ISO-30042 2019). This document defines the TBX framework, expressed in XML markup language, used in analysis, descriptive representation and dissemination of structured terminological data. Moreover, the TBXinfo.net [31] website provides many examples which may prove useful to those who wish to structure their data in TBX format. The main purpose of the TBX framework is to ensure that lexical data can be used in different software. For istance, in the translation process with CAT tools it is the most commonly used format for terminology management systems (Bowker and Fisher 2010).

The structure of the TBX format is structured in two interacting components: 1) a basic structure that reflects the TMF meta-model (see section 3.1); and 2) a formalism aiming to define modules with a list of data categories. There are two XML styles that can be used to represent terminology data: Data Category as Attribute (DCA) and Data Category as Tag (DCT).[32] The combination of these two components defines a particular "dialect", that is, an XML markup language conforming to TBX. Dialects may differ in terms of permitted data categories and meta-model levels where such categories may be entered. On the TBXinfo.net website, three public dialects are described and recommended for the exchange of terminology: 1) TBX-Core, 2) TBX-Min, 3) TBX-Basic.[33] These dialects provide a restricted set of data categories. The standard therefore allows the creation of "private dialects" for anyone wishing to represent terminological data categories that are not included in the previously stated modules.

31. `https://www.tbxinfo.net`

32. The value of the data category is given by the content of the XML element and the two styles are isomorphic: that is, they can be converted in both directions by an algorithm without loss of information. For more information, see section 6 of the ISO standard or `https://www.tbxinfo.net/dca-v-dct/`

33. `https://www.tbxinfo.net/tbx-dialects/`

3.3.1 TBX-TriMED dialect

Considering the TriMED data categories, we therefore decided to create a new dialect. For already documented data categories, the standard specifies that public modules must be referred to, while for the other categories, a new specific module must be created. The TriMED dialect therefore includes the following modules:

- *Core* which contains the following data category: /term/, /date/ and /note/.
- *Min* which contains (among others) the following data category: /part of speech/and /subject field/.
- *Basic* which contains (among others) the following data category: /context/, /definition/, /cross reference/, /grammatical gender/, /source/, /responsibility/, /transaction type/ and /external cross reference/.
- *Trimed* which contains the following data category: /superordinate concept/, /subordinate concept/, /subdomain/, /concept identifier/, /language identifier/, /identifier/, /grammatical number/, /derivative/, /pronunciation/, /etymology/, /orthographic variant/, /abbreviation/, /full form/, /acronym/, /semic analysis/, /synonym/, /hyponym/, /hypernym/, /phraseological unit/, /collocation/, /register/, /common name/, /scientific name/, /ICPC2 code/, /ICD10 code/, /SNOMED CT term/, /MeSh term/, /concept relation/ and /conceptual sphere/.

To comply with the standard, the formal definition of the *Trimed* module was written in prose using the TBX Module Description (TBXMD) formalism. In Figure 3, we provide a partial list of the data categories contained in our module with all the relevant information following international standards which include: the identifier, PID, classification, value

Data Category Module Definition

Name: Trimed

Namespace: http://shiny.dei.unipd.it/TriMED/data_category_repository/

Description: The Trimed Module is a data category module collecting data categories selected from 1) DatCatInfo (http://www.datcatinfo.net/#/) and 2) TriMED Data Category Repository.

Data Categories:

Name	PID	Classification	Value(s)	Level (TMF)
superordinateConcept	http://www.datcatinfo.net/datcat/DC-495	<descrip>		conceptEntry
subordinateConcept	http://www.datcatinfo.net/datcat/DC-490	<descrip>		conceptEntry
subDomain	http://purl.org/trimed/dcr/dc/dc-7	<descrip>		conceptEntry
conceptIdentifier	http://www.datcatinfo.net/datcat/DC-139	<descrip>		conceptEntry
languageIdent	http://www.datcatinfo.net/datcat/DC-279	<descrip>		langSec
identifier	http://www.datcatinfo.net/datcat/DC-3894	<descrip>		termSec
grammaticalNumber	http://www.datcatinfo.net/datcat/DC-251	<termNote>	singular, plural, dual, mass, otherNumber	termSec
derivative	http://www.datcatinfo.net/datcat/DC-4611	<termNote>		termSec
pronunciation	http://www.datcatinfo.net/datcat/DC-407	<termNote>		termSec

Figure 3: TBX Module Description of TriMED

and TMF level.[34]

Once the components of the TBX-TriMED dialect and the data categories for each module were defined, we provide an instance excerpt of our dialect in the appendix of this article. In the root element <*tbx*>, the following information are specified: 1) the TBX-Core,[35] TBX-Min, TBX-Basic and TBX-TriMED namespaces; 2) the value of the type attribute, that is the name of the TBX dialect; 3) the style of the instance, that is the DCT format; and 4) the working language (xml: lang) of the document. This last attribute should not be confused with the <*langSec*> element: the definition of the working language in the root element of the document indicates the language in which the

34. The TriMED dialect is currently under review. A partial dialect definition package is available at: `https://github.com/trimed-dialect-2020/TriMED`

35. That is *urn:iso:std:iso:30042:ed-2* which has to be used as the default namespace for TBX document instances of all dialects.

TBX file is written (for example the tags). Whereas, the <*langSec*> specifies the language described in that specific section of the TBX file (for example the tag content).

The instance illustrates how the TriMED terminological data are organized. It adopts the TMF hierarchical meta-model in the following order: i) <*conceptEntry*>, ii) <*langSec*>, and iii) <*termSec*> are defined. At the concept level, represented by its identifier (id=A2), the data categories provided are: /conceptual sphere/, /cross reference/, /concept relation/, /transaction/, /responsibility/, /date/, /note/, /superordinate concept/, /subordinate concept/,/subject field/ and /subdomain/. The same concept is then expressed in three language sections, each containing the terms designating the concept in English, French and Italian. At the language level, we provide the information about the /definition/, /external cross reference/, /source/, /note/ and /semic analysis/ of the concept. Each <*langSec*> contains a <*termSec*> with the term designating the concept and other data categories, such as /part of speech/, /grammatical gender/ and /derivative forms/ etc. For each data category of the TBX instance, the corresponding module is specified in the XML language.

3.3.2 Element position

Finally, the standard sets the position constraints of the data categories for each level of the TMF meta-model. In general, more generic data categories have a certain degree of freedom in terms of positioning. For example, the /definition/ can be placed both at the concept and at the language level, while the descriptive information on the /grammatical number/ of the term must only be placed at the term section level. In Figure 4, we show the TriMED data categories assigned for each TMF hierarchical level (concept, language and term) and the TBX modules to which the categories belong to (core, min, basic and trimed). Categories were positioned following standard indications, where constraint

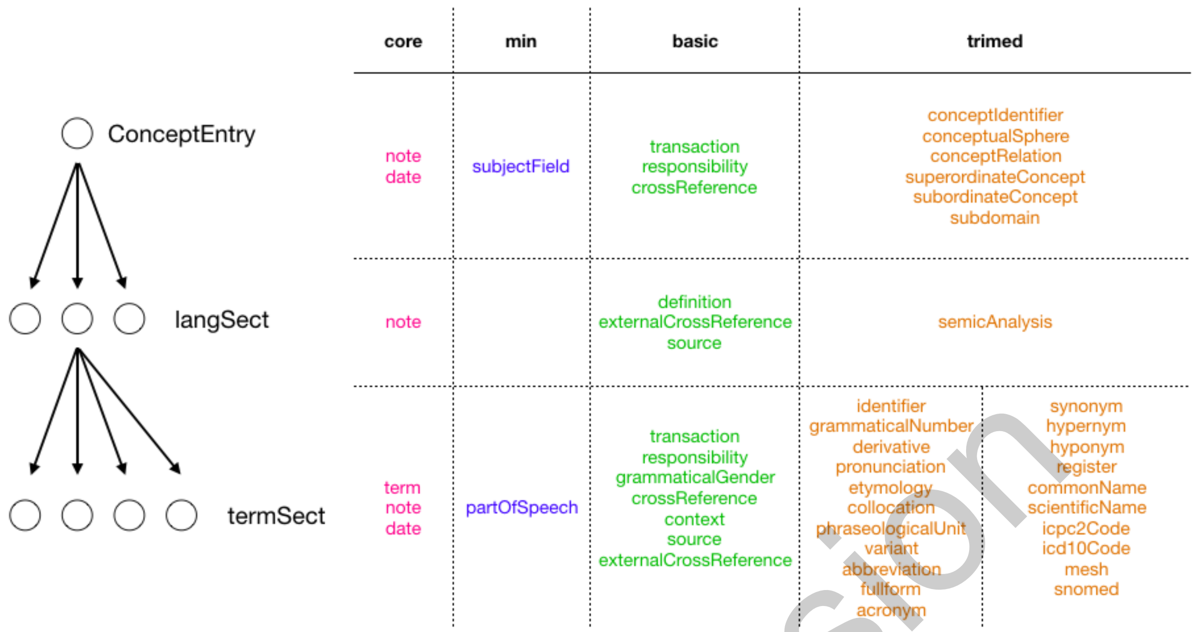

Figure 4: Element Position in TriMED

levels are explicit. For those categories without restrictions we deliberately decided their positioning level. For example, we assigned the /definition/ data category at the language level, and not at the concept level as in the TermScience resource (Khayari et al. 2006). This avoids assigning a linguistic identity to a concept whose meaning should be equivalent for all the languages analyzed in the terminology entry. Had the definition been placed at the concept level, an arbitrary language would have to be chosen for its representation. Instead, by placing the definition at the language level, it can be expressed in each of the object languages of the resource (English, Italian and French). In this way, the concept (and consequently the meaning of a term) remains a *super partes* entity unrelated to any linguistic connotation. Finally, such decision makes it possible to insert, in the level below the language section, all the terms (one per term section) that correspond exactly to that definition. In this way, a concept can be expressed in a language that uses equivalent terms conveying the same meaning.

4 Web Application

The TriMED Web application was implemented using the source code provided by Vezzani, Di Nunzio, and Henrot 2018 with the Shiny R package (Chang 2015), available online.[36] From the homepage, users can select one of the three consultation categories (patient, physician and translator) and then access the three interfaces which provide different subset of linguistic information depending on user identification.

4.1 Patient

To manage the diastratic variation within expert to non-expert dialogues and to facilitate the comprehension of medical information, patients can look up a scientific term and then consult the corresponding common name and its definition, if available. For example, in French the scientific term *rougeole* (measles) is commonly referred to as *première maladie*.

A patient looking for information on a popular term starts the research by selecting a language (English, French, or Italian) and then typing the term in the search box (Fig. 5). The system automatically filters words, character by character, and shows the possible alternatives in the box. After careful analysis, we decided to display both the popular and the technical term during the search to give users immediate feedback. When the user selects the term, the definition and the information on terminological variation are displayed on the page.[37]

4.2 Physician

As we mentioned above, interoperability of resources is a fundamental requirement when adopting a FAIR approach to terminology. In this context, one of TriMED's most ambi-

36. `http://purl.org/TriMED`

37. The direct link to the patient interface is the following: `http://purl.org/TriMED/patient`

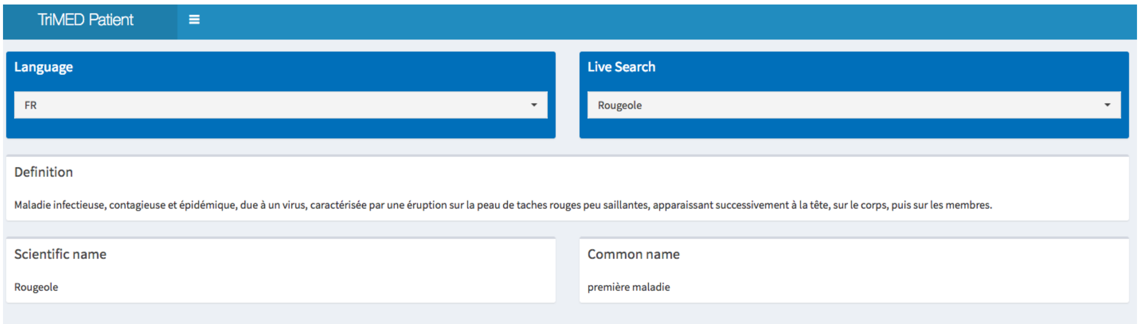

Figure 5: Patient interface

tious goals is to provide a system that collects medical information from multiple sources, through a single access point. Hence, when consulting the TriMED database, physicians can look up a technical term in one of the three working languages. They can therefore access standard medical terminologies, ontologies and classification systems, such as MeSH terms and SNOMED CT concepts.

In (Fig. 6) we show the physician's interface. First, the user can select the working language and then, as in the case of the patient interface, s/he can enter the term in the search box: the system will automatically filter words and characters showing the user possible options. The system will automatically come up with the link to other external resources and provide direct access to the same term and/or combination of terms that are correlated to the main search. In the example provided in (Fig. 6), the expert looking for the term "Symptom" will have direct access to other related medical information such as "Symptom Assessment", "Symptom Cluster" or "Symptom Evaluation" provided by the Medical Subject Headings (MeSH) controlled vocabulary.[38]

38. The direct link to the physician interface is the following: `http://purl.org/TriMED/physician`

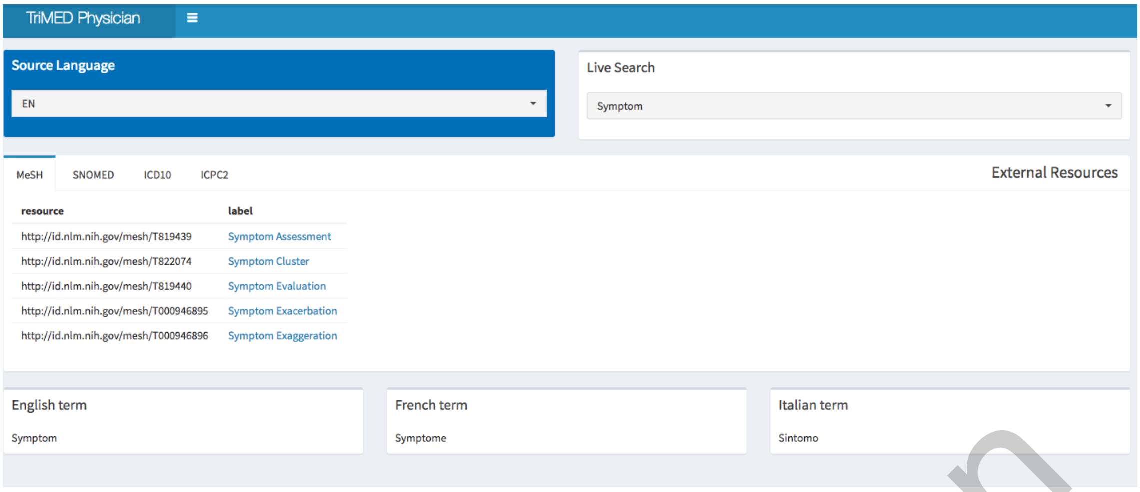

Figure 6: Physician interface

4.3 Translator

Many studies have focused on the complexity of medical translation from different perspectives (see, among others: Jammal 1999; Rouleau 2003; Karwacka 2014; Peters, Qian, and Ding 2018; Alarcón-Navío, López-Rodríguez, and Tercedor-Sánchez 2016; Nisbeth Brøgger 2017). With the TriMED database, we strive to provide a useful tool help technical-scientific translators in their specialized translation activity. The record model for the translator consultation embodies 42 different data categories for both the source and target terms offering a complete framework on syntactic, semantic and phraseological behaviour of terms. As shown in Fig. 7, the record is organized into six information panels providing different data categories:

1. Formal features: /part of speech/, /grammatical gender/, /grammatical number/, /pronunciation/, /etymology/ and /derivative forms/;

2. Semantics: /definition/, /external cross reference/, /source/, /note/ and /semic analysis/;

3. Variation: /orthographic variant/, /acronym/, /full form/, /abbreviation/, /syn-

onym/, /hypernym/, /hyponym/, /cross reference/, /common name/ and /scientific name/;

4. Usage: /context/, /cross reference/, /source/, /collocation/, /phraseological unit/ and /register/;

5. Administrative information about the concept: /concept ID/, /date/, /note/, /responsibility/, /transaction/, /conceptual sphere/, /domain/, /subdomain/, /concept relation/, /cross reference/, /superordinate concept/ and /subordinate concept/;

6. Administrative information about the term: /term ID/, /transaction/, /responsibility/, /date/, /note/.

The user can access all linguistic features of the multilingual version of every record in TriMED. For example, by examining the terms shown in Fig. 7, that is *rougeole* for French and *morbillo* for Italian, the translator can consult, in the "Usage" panel, all the collocations (i.e. preferential and frequent co-occurrences) for these two terms: *attraper la rougeole, être atteint de rougeole* and *prendere il morbillo, contrarre il morbillo.* Moreover, the translator can download the standard TBX format of the record, compatible with any Computer Assisted Translation software that supports the standard. The left sidebar of the interface enables the translator to select one of three visualization options: Search, XML, and Download. The translator must first choose at least one source language in the Search visualization to activate the search box. After selecting a term, the interface will show the data categories grouped into six panels. If the user selects a target language, the XML and the Download visualization areas will display 1) the TBX version of the

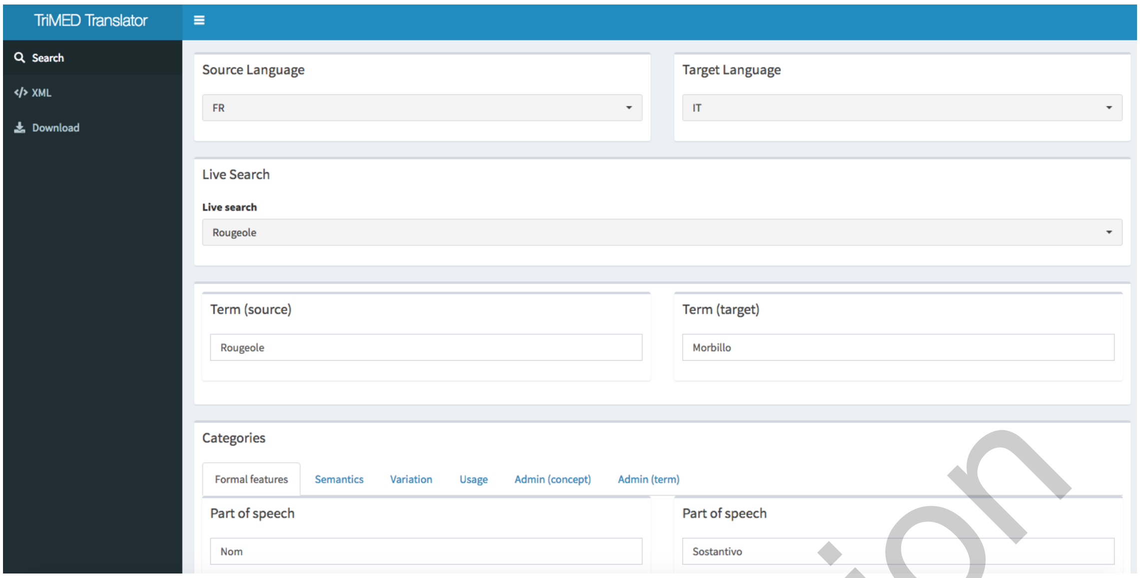

Figure 7: Translator interface

multilingual record and 2) a Download button to download the TBX record.[39]

4.4 Dataset

Currently, the database collects a total of 789 multilingual terminological records. Record compilation work has been running since 2018 and, at present, the following number of completed standardized records is available per language: 388 for English, 160 for French and 241 for Italian. Each terminological record has been compiled thanks to the collaboration of different student groups as part of the Master degree course in Modern Languages for International Communication and Cooperation at the University of Padua.[40] In particular, we worked on several projects aimed at the study of medical terminology for the process of specialized translation (Vezzani and Di Nunzio 2019b) and query reformulation for systematic medical reviews (Di Nunzio and Vezzani 2018; Vezzani and Di Nunzio

39. The direct link to the translator interface is the following: `http://purl.org/TriMED/translator`

40. `https://elearning.unipd.it/scienzeumane/course/view.php?id=4226`

2019a). We are currently working on a qualitative content revision of the records and are collecting more terminological data in order to balance the number of records per language.

5 Conclusions

In this paper, we have focused on two different aspects involving the content and the structure of terminology in the context of terminology standardization. We have defined 1) the notion of "semantic standardization" as all the initiatives carried out in order to standardize the meaning of technical terms and create a common shared consensus on the use of terms; and 2) the notion of "structural standardization" as all the activities aimed at the standardizing of the structure of language resources. The latter relates strictly to the FAIR principles of the European Open Science Cloud (ESOC) that ensure interoperability and reusability of terminological infrastructure.

In particular, we have presented a general methodology for structural standardization of terminological resources according to three of the latest standards provided by the ISO TC/37, sub-committee 3, dealing with the management of terminology resources. We have applied this methodology on the TriMED multilingual termbase and we have focused on 1) the definition of the resource structure according to Terminological Markup Framework (TMF) meta-model; 2) the description of its data category specifications through the implementation of the TriMED Data Category Repository, and, to our knowledge, this is the first DCR implementation that complies with the latest ISO standards; 3) the implementation of the resource in the TermBase eXchange (TBX) format; and 4) the realization of the Web application to access the data stored in the termbase. Despite the conceptual orientation of the proposed methodology, we have showed how this structure can be used to represent the natural phenomena of terminological variation.

In terms of future work, we are working on the population of the TriMED database and, for this purpose, we are collecting terminological data on the Covid-19 outbreak. In relation to the design of our web application, we are working on a localization system that provides access to the resource in the three working languages. We also envisage the possibility of having external users that contribute in the compilation of the missing terminological records, adding even new ones. Finally, we are reviewing the full specification of the TriMED dialect definition package to be submitted for approval to the TBX.info website.

6 Funding information

This work was partially supported by the ExaMode Project, as a part of the European Union Horizon 2020 Program under Grant 825292.

References

Alarcón-Navío, Esperanza, Clara López-Rodríguez, and Maribel Tercedor-Sánchez. 2016. "Variation dénominative et familiarité en tant que source d'incertitude en traduction médicale." *Meta: journal des traducteurs/Meta: Translators' Journal* 61 (1): 117–144.

Association, Management, et al. 2015. *Standards and Standardization: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications.* IGI Global.

Benson, Tim. 2012. *Principles of Health Interoperability HL7 and SNOMED.* Springer Science & Business Media.

Bowker, Lynne, and Des Fisher. 2010. "Computer-aided Translation." *Handbook of Translation Studies* 1:60–65.

Bowker, Lynne, and Shane Hawkins. 2006. "Variation in the Organization of Medical Terms: Exploring some Motivations for Term Choice." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 12 (1): 79–110.

Bowman, Sue E. 2005. "Coordination of SNOMED-CT and ICD-10: Getting the most out of Electronic Health Record Systems." *Coordination of SNOMED-CT and ICD-10: Getting the Most out of Electronic Health Record Systems/AHIMA, American Health Information Management Association.*

Broeder, Daan, Ineke Schuurman, and Menzo Windhouwer. 2014. "Experiences with the ISOcat Data Category Registry." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* 4565–4568. Reykjavik, Iceland: European Language Resources Association (ELRA), May. http://www.lrec-conf.org/proceedings/lrec2014/pdf/153_Paper.pdf.

Cabré Castellví, Maria Teresa. 1997. "Standardization and Interference in Terminology." *The Changing Scene in World Languages: Issues and Challenges* 9:49.

———. 2003. "Theories of Terminology: their Description, Prescription and Explanation." *Terminology* 9 (2): 163–199.

Cardillo, Elena, Josepph Roumier, Marc Jamoulle, Robert Vander Stichele, et al. 2013. "Using ISO and Semantic Web standards for Creating a Multilingual Medical Interface Terminology: a Use Case for Hearth Failure." *Terminology and Artificial Intelligence TIA 2013.*

Cardillo, Elena, Luciano Serafini, and Andrei Tamilin. 2009. "A Lexical-ontological Resource for Consumer Healthcare." In *International Conference on Electronic Healthcare,* 131–138. Springer.

Castro, Cesar M, Clifford Wilson, Frances Wang, and Dean Schillinger. 2007. "Babel Babble: Physicians' use of Unclarified Medical Jargon with Patients." *American Journal of Health Behavior* 31 (1): S85–S95.

Chang, Winston. 2015. *Shiny: Web Application Framework for R.* R package version 0.11. `http://CRAN.R-project.org/package=shiny`.

Corti, Louise, Veerle Van den Eynden, Libby Bishop, and Matthew Woollard. 2019. *Managing and Sharing Research Data: a Guide to Good Practice.* SAGE Publications Limited.

Coseriu, Eugenio. 1969. *Einführung in die strukturelle Linguistik.* Autorisierte Nachschrift besorgt von Gunter Narr und Rudolf Windisch.

Da Silva Santos, Luiz Bonino, Mark D Wilkinson, Arnold Kuzniar, Rajaram Kaliyaperumal, Mark Thompson, Michel Dumontier, and Kees Burger. 2016. "FAIR Data Points Supporting Big Data Interoperability." *Enterprise Interoperability in the Digitized and Networked Factory of the Future. ISTE, London:* 270–279.

De Quirós, Fernán González Bernaldo, Carlos Otero, and Daniel Luna. 2018. "Terminology Services: Standard Terminologies to Control Health Vocabulary." *Yearbook of Medical Informatics* 27 (01): 227–233.

Delavigne, Valérie. 2017. "Term Usage and Socioterminological Variation." In *Multiple Perspectives on Terminological Variation.* 18:31. John Benjamins Publishing Company.

Deuster, Lindsay, Stephanie Christopher, Jodi Donovan, and Michael Farrell. 2008. "A Method to Quantify Residents' Jargon use during Counseling of Standardized patients about Cancer Screening." *Journal of General Internal Medicine* 23 (12): 1947.

Di Nunzio, Giorgio Maria, and Federica Vezzani. 2018. "Using R Markdown for Replicable Experiments in Evidence Based Medicine." In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings,* edited by Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian-Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato, and Nicola Ferro, 11018:28–39. Lecture Notes in Computer Science. Springer. doi:10.1007/978-3-319-98932-7_3. https://doi.org/10.1007/978-3-319-98932-7%5C_3.

Faber, Pamela. 2015. "Frames as a Framework for Terminology." *Handbook of terminology* 1 (14): 14–33.

Felber, Helmut. 1980. "International Standardization of Terminology: Theoretical and Methodological Aspects." *International Journal of the Sociology of Language* 1980 (23): 65–80.

Freixa, Judit. 2006. "Causes of Denominative Variation in Terminology: a Typology Proposal." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 12 (1): 51–77.

Galinski, Christian, and Wolfgang Nedobity. 1988. "Special Languages, Terminology Planning and Standardization." In *Standardization of Technical Terminology: Principles and Practices (Second Volume).* ASTM International.

Ingenerf, Josef, Jörg Reiner, and Bettina Seik. 2001. "Standardized Terminological Services Enabling Semantic Interoperability between Distributed and Heterogeneous Systems." *International Journal of Medical Informatics* 64 (2-3): 223–240.

ISO-12620. 2019. *Management of Terminology Resources – Data Category Specifications.* Standard. Geneva, CH: International Organization for Standardization, March.

ISO-16642. 2017a. *Computer Applications in Terminology – Terminological Markup Framework.* Standard. Geneva, CH: International Organization for Standardization, November.

ISO-171171. 2018. *Health informatics – Terminological resources – Part 1: Characteristics.* Standard. Geneva, CH: International Organization for Standardization, April.

ISO-30042. 2019. *Management of terminology resources – TermBase eXchange (TBX).* Standard. Geneva, CH: International Organization for Standardization, April.

ISO-16642. 2017b. *ISO 16642. Computers Applications in Terminology – Terminological Markup Framework.* 2nd ed. Geneva. International Organization for Standardization.

Jammal, Amal. 1999. "Une méthodologie de la traduction médicale." *Meta: journal des traducteurs/Meta: Translators' Journal* 44 (2): 217–237.

Jucks, Regina, and Rainer Bromme. 2007. *Choice of Words in Doctor-patient Communication: an Analysis of Health-related Internet Sites.* Health Commun, 21(3):267-77.

Kachlik, David, Vaclav Baca, Ivana Bozdechova, Pavel Cech, and Vladimir Musil. 2008. "Anatomical Terminology and Nomenclature: Past, Present and Highlights." *Surgical and Radiologic Anatomy* 30 (6): 459.

Karwacka, Wioleta. 2014. "Quality Assurance in Medical Translation." *The Journal of Specialised Translation* 21:19–34.

Kemps-Snijders, Marc, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2008. "ISOcat: Corralling Data Categories in the Wild." In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).* Marrakech, Morocco: European Language Resources Association (ELRA), May. http://www.lrec-conf.org/proceedings/lrec2008/pdf/222_paper.pdf.

Khayari, Majid, Stéphane Schneider, Isabelle Kramer, Laurent Romary, et al. 2006. "Unification of Multi-lingual Scientific Terminological Resources Using the ISO 16642 Standard. The TermSciences Initiative." *arXiv preprint cs/0604027.*

L'Homme, Marie-Claude. 2004. *La terminologie: principes et techniques.* Pum.

Lee, Dennis, Nicolette de Keizer, Francis Lau, and Ronald Cornet. 2013. "Literature Review of SNOMED CT Use." *Journal of the American Medical Informatics Association* 21 (e1): e11–e19.

León-Araúz, Pilar, Arianne Reimerink, and Pamela Faber. 2019. "EcoLexicon and By-products: Integrating and Reusing Terminological Resources." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 25 (2): 222–258.

Lindberg, Donald AB, Betsy L Humphreys, and Alexa T McCray. 1993. "The Unified Medical Language System." *Yearbook of Medical Informatics* 2 (01): 41–51.

Lipscomb, Carolyn E. 2000. "Medical Subject Headings (MeSH)." *Bulletin of the Medical Library Association* 88 (3): 265.

Llanos, Leonardo Campillos, Dhouha Bouamor, Pierre Zweigenbaum, and Sophie Rosset. 2016. "Managing Linguistic and Terminological Variation in a Medical Dialogue System." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 3167–3173.

Magris, Marella, Maria Teresa Musacchio, Lorenza Rega, and Federica Scarpa. 2002. *Manuale di terminologia: aspetti teorici, metodologici e applicativi.* Hoepli.

Melby, Alan. 1995. "Implementing the Terminology Interchange Format." In *Standardizing and Harmonizing Terminology: Theory and Practice.* ASTM International.

———. 2015. "TBX: a Terminology Exchange Format for the Translation and Localization Industry." *201), Handbook of Terminology:* 393–424.

Melby, Alan, Klaus-Dirk Schmitz, and Sue Ellen Wright. 2001. "Terminology Interchange." *Handbook of Terminology Management: Volume 2: Application-Oriented Terminology Management:* 613.

Müller, Kirill, and Hadley Wickham. 2019. *Tibble: Simple Data Frames.* R package version 2.1.3. https://CRAN.R-project.org/package=tibble.

Nisbeth Brøgger, Matilde. 2017. "When Translation Competence is not Enough: a Focus Group Study of Medical Translators." *Meta: journal des traducteurs/Meta: Translators' Journal* 62 (2): 396–414.

Palmer, Carole, Nicholas M Weber, Allen H Renear, and Trevor Muñoz. 2013. "Foundations of Data Curation: the Pedagogy and Practice of Purposeful Work with Research Data."

Pavel, Silvia, and Diane Nolet. 2001. "Précis de terminologie/The Handbook of Terminology." *Adapted into English by Christine Leonhardt. Hull: Public Works and Government Services of Canada.*

Pecman, Mojca. 2018. *Langue et construction de connaisSENSes. Energie lexico-discursive et potentiel sémiotique des sciences. Préface de Marie-Claude L'Homme.*

Peters, Pam, Yan Qian, and Jun Ding. 2018. "Translating Medical Terminology and Bilingual Terminography." *Lexicography* 3 (2): 99–113.

Picht, Heribert, and Klaus-Dirk Schmitz. 2001. *Terminoligie und Wissensordnung: Ausgewählte Schriften aus dem Gesamtwerk von Eugen Wüster.* TermNet publisher.

Picton, Aurélie, and Pascaline Dury. 2017. "Diastratic Variation in Language for Specific Purposes." In *Multiple Perspectives on Terminological Variation.* 18:57. John Benjamins Publishing Company.

Pitar, Mariana, et al. 2011. "La fiche terminologique–expansion et applications." *Buletinul Stiintific al Universitatii Politehnica din Timisoara, Seria Limbi Moderne,* no. 10: 70–83.

Rastier, François. 2009. *Sémantique interprétative.* Presses universitaires de France.

Romary, Laurent. 2001a. "An Abstract Model for the Representation of Multilingual Terminological Data: TMF-Terminological Markup Framework." In *TAMA 2001.*

———. 2001b. "Un modèle abstrait pour la représentation de terminologies multilingues informatisées."

Romary, Laurent, Isabelle Kramer, Susanne Salmon-Alt, and Joseph Roumier. 2006. *Gestion de données terminologiques: principes, modèles, méthodes.*

Rouleau, Maurice. 2003. *La terminologie médicale et ses problèmes.* Tribuna, Vol. IV, n. 12.

Roumier, Joseph, Robert Vander Stichele, Laurent Romary, and Elena Cardillo. 2011. "Approach to the Creation of a Multilingual, Medical Interface Terminology." In *Workshop Proceedings of the 9th International Conference on Terminology and Artificial Intelligence,* 13–15. INALCO.

Sager, Juan C. 1998. "Terminology: Standardization." *& "Terminology: theory". In Mona Baker (Ed.) Routledge Encyclopedia of Translation Studies. GB: Routledge.*

Sánchez, Maribel Tercedor, and Juan Antonio Prieto Velasco. 2013. "Las barreras en la comunicación médico-paciente: el proyecto VARIMED." In *Translating culture,* 593–606. Editorial Comares.

Schmitz, Klaus-Dirk. 2012. "Using International Standards for Terminology Exchange." *Terminologija* 19 (2012): 33–38.

Strehlow, Richard Alan. 1988. *Standardization of Technical Terminology: Principles and Practices (second volume).* Vol. 991. ASTM International.

Stührenberg, Maik. 2012. "The TEI and Current Standards for Structuring Linguistic Data. An Overview." *Journal of the Text Encoding Initiative,* no. 3.

Temmerman, Rita. 2000. *Towards New Ways of Terminology Description: The Sociocognitive-approach.* Vol. 3. John Benjamins Publishing.

Van Campenhoudt, Marc. 2017. "Standardised Modelling and Interchange of Lexical Data in Specialised Language." *Revue française de linguistique appliquée* 22 (1): 41–60.

Vezzani, Federica, and Giorgio Maria Di Nunzio. 2019a. "Computational Terminology in eHealth." In *Digital Libraries: Supporting Open Science - 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Pisa, Italy, January 31 - February 1, 2019, Proceedings,* edited by Paolo Manghi, Leonardo Candela, and Gianmaria Silvello, 988:72–85. Communications in Computer and Information Science. Springer. doi:10.1007/978-3-030-11226-4_6. https://doi.org/10.1007/978-3-030-11226-4%5C_6.

———. 2019b. "On the Use of Terminological Records in Specialised Translation." In *Proceedings of Associazione per l'Informatica Umanistica e le Culture Digitali (AIUCD).* Udine, Italy.

Vezzani, Federica, Giorgio Maria Di Nunzio, and Geneviève Henrot. 2018. "TriMED: A Multilingual Terminological Database." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*

Warnier, Maxime, Joseph Roumier, Marc Jamoulle, Elena Cardillo, Robert Vander Stichele, and Laurent Romary. 2012. "Publishing a Multilingual Medical Terminology According to Terminology Standards and Linked Data Principles." In *SABRE Conference 2012.*

Wermuth, Cornelia, and Heidi Verplaetse. 2019. "Medical Terminology in the Western World. Current Situation." [in eng]. In *Handbook of Terminology, Volume 2 [HoT2]: Terminology in the Arab world,* 2:84–108. John Benjamins; Amsterdam. ISBN: 9789027201980. $$Uhttps://lirias.kuleuven.be/retrieve/531149$

$D2018_Benjamins%20Offprint_HoT2_Ch4_Medical%20Terminology%20in%20the%20Western%20World.pdf%20[freely%20available].

Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific data* 3.

Windhouwer, Menzo, and Ineke Schuurman. 2014. "Linguistic Resources and Cats: How to Use ISOcat, RELcat and SCHEMAcat." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* 3806–3810. Reykjavik, Iceland: European Language Resources Association (ELRA), May. http://www.lrec-conf.org/proceedings/lrec2014/pdf/396_Paper.pdf.

Wright, Sue Ellen. 2001. "Data Categories for Terminology Management." *Handbook of Terminology Management: Volume 2: Application-Oriented Terminology Management:* 552.

Wüster, Eugen. 1971. "Principles of Special Language Standardization." *Muttersprache* 81 (5): 289–295.

———. 1979. "Introduction to the General Theory of Terminology and Terminological Lexicography." *Springer, Vienna, part* 1:145.

Zeng, Qing T, and Tony Tse. 2006. "Exploring and Developing Consumer Health Vocabularies." *Journal of the American Medical Informatics Association* 13 (1): 24–29.

Appendix


```xml
<?xml version="1.0" encoding="UTF-8"?>

<tbx type="TBX-Trimed" style="dct" xml:lang="en" xmlns="urn:iso:std:iso:30042:ed-2"
xmlns:min="http://www.tbxinfo.net/ns/min" xmlns:basic="http://www.tbxinfo.net/ns/basic"
xmlns:trimed="https://github.com/trimed-dialect/TriMED">
        <tbxHeader>
        <fileDesc>
            <sourceDesc>
                <p>A sample of the TriMED termbase consisting of one concept entry</p>
            </sourceDesc>
        </fileDesc>
    </tbxHeader>
    <text>
        <body>
            <conceptEntry>
             <trimed:conceptIdentifier>A2</trimed:conceptIdentifier>
                <trimed:conceptualSphere>A</trimed:conceptualSphere>
                <basic:crossReference></basic:crossReference>
                <trimed:conceptRelation></trimed:conceptRelation>
                <basic:transaction>origination</basic:transaction>
                <basic:responsibility>FV</basic:responsibility>
                <date>04/04/2020</date>
                <note></note>
                <trimed:superordinateConcept></trimed:superordinateConcept>
                <trimed:subordinateConcept></trimed:subordinateConcept>
                <min:subjectField>medicine</min:subjectField>
                <trimed:subDomain></trimed:subDomain>

                <langSec xml:lang="en">
                    <basic:definition>The act or process of identifying or determining the nature
                        and cause of a disease or injury through evaluation of patient history,
                        examination, and review of laboratory data.</basic:definition>
                    <basic:externalCrossReference>https://www.thefreedictionary.com/diagnosis</basic:externalCrossReference>
```

<basic:source></basic:source>

<note></note>

<trimed:semicAnalysis>/process/ /identification/ /nature/ /cause/ /disease/</trimed:semicAnalysis>

<termSec>

<term>Diagnosis</term>

<trimed:identifier>TPLT18167</trimed:identifier>

<basic:transaction>origination</basic:transaction>

<basic:responsibility>Student TPLT</basic:responsibility>

<date>2018</date>

<note></note>

<min:partOfSpeech>noun</min:partOfSpeech>

<basic:grammaticalGender></basic:grammaticalGender>

<trimed:grammaticalNumber></trimed:grammaticalNumber>

<trimed:pronunciation></trimed:pronunciation>

<trimed:etymology></trimed:etymology>

<trimed:variant>plural: diagnoses</trimed:variant>

<trimed:acronym></trimed:acronym>

<trimed:fullForm></trimed:fullForm>

<trimed:abbreviation></trimed:abbreviation>

<trimed:derivative>to diagnose, to diagnosticate, diagnostic, diagnostically, diagnostician</trimed:derivative>

<trimed:collocation>accurate diagnosis, age of diagnosis, clinical diagnosis, correct diagnosis,definitive diagnosis, early diagnosis, differential diagnosis, diagnosis of, difficulty of diagnosis</trimed:collocation>

<trimed:phraseologicalUnit></trimed:phraseologicalUnit>

<trimed:synonym>analysis, examination, investigation, conclusion, interpretation</trimed:synonym>

<trimed:hypernym></trimed:hypernym>

<trimed:hyponym></trimed:hyponym>

<basic:crossReference></basic:crossReference>

```
<basic:context>Consequently, people with suspected hypertension have been
    required to undergo repeated measurements of their clinic BP on repeated
    clinic visits to confirm or refute the diagnosis of hypertension</basic:context>
<basic:externalCrossReference>https://www.nice.org.uk/guidance/ng136/
evidence/full−guideline−pdf−6898565198</basic:externalCrossReference>
<basic:source></basic:source>

<trimed:register>specialized</trimed:register>
<trimed:commonName>Diagnosis</trimed:commonName>
<trimed:scientificName>Diagnosis</trimed:scientificName>
<trimed:mesh>https://meshb.nlm.nih.gov/record/ui?ui=D003933</trimed:mesh>
<trimed:snomed>https://browser.ihtsdotools.org/?perspective=full&
amp;conceptId1=439401001&amp;edition=MAIN
2020−03−09&amp;release=&amp;languages=en</trimed:snomed>
<trimed:icpc2Code></trimed:icpc2Code>
<trimed:icd10Code></trimed:icd10Code>

</termSec>
</langSec>

<langSec xml:lang="fr">
<basic:definition>Connaissance que l'on peut avoir d'une maladie en observant
    les signes de celle−ci; art du diagnostic.</basic:definition>
<basic:externalCrossReference>http://stella.atilf.fr/Dendien/scripts/tlfiv5/
advanced.exe?33;s=1437258285;</basic:externalCrossReference>
<basic:source></basic:source>
<note></note>
<trimed:semicAnalysis>/etude/maladie/symptome/medecine/</trimed:semicAnalysis>

<termSec>
    <term>Diagnose</term>
    <trimed:identifier>TSF18068</trimed:identifier>
    <basic:transaction>Origine</basic:transaction>
    <basic:responsibility>etudiants TRADFR</basic:responsibility>
    <date>2018</date>
```

<note></note>

<min:partOfSpeech>nom</min:partOfSpeech>

<basic:grammaticalGender>f.</basic:grammaticalGender>

<trimed:grammaticalNumber></trimed:grammaticalNumber>

<trimed:pronunciation>\djagno:z\</trimed:pronunciation>

<trimed:etymology>Empr. au gr. proprement action de discerner

spec. terme de med. diagnose, diagnostic.</trimed:etymology>

<trimed:variant></trimed:variant>

<trimed:acronym></trimed:acronym>

<trimed:fullForm></trimed:fullForm>

<trimed:abbreviation></trimed:abbreviation>

<trimed:derivative>diagnostic, diagnostique, diagnostiquer</trimed:derivative>

<trimed:collocation>faire une diagnose, effectuer une diagnose</trimed:collocation>

<trimed:phraseologicalUnit></trimed:phraseologicalUnit>

<trimed:synonym>diagnostic</trimed:synonym>

<trimed:hypernym></trimed:hypernym>

<trimed:hyponym></trimed:hyponym>

<basic:crossReference></basic:crossReference>

<basic:context>Les courbes de temperature (...) ne sont pas

suffisamment typiques pour qu'elles puissent servir a la diagnose</basic:context>

<basic:externalCrossReference>http://stella.atilf.fr/Dendien/

scripts/tlfiv5/advanced.exe?33;s=1437258285;</basic:externalCrossReference>

<basic:source></basic:source>

<trimed:register>specialise</trimed:register>

<trimed:commonName>Diagnose</trimed:commonName>

<trimed:scientificName>Diagnose</trimed:scientificName>

<trimed:mesh>https://meshb.nlm.nih.gov/record/ui?ui=D003933</trimed:mesh>

<trimed:snomed>https://browser.ihtsdotools.org/?

perspective=full&conceptId1=439401001&edition=MAIN/

2020−03−09&release=&languages=en</trimed:snomed>

<trimed:icpc2Code></trimed:icpc2Code>

<trimed:icd10Code></trimed:icd10Code>

</termSec>

</langSec>

<langSec xml:lang="it">

<basic:definition>Procedura consistente nell'interpretazione di segni e sintomi raccolti quali manifestazioni di un processo patologico in atto o pregresso<basic:definition>

<basic:externalCrossReference>http://www.treccani.it/enciclopedia/diagnosi_%28Dizionario-di-Medicina%29/</basic:externalCrossReference>

<basic:source></basic:source>

<note></note>

<trimed:semicAnalysis>/procedura/ /interpretazione/ /sintomi/ /malattia/</trimed:semicAnalysis>

<termSec>

<term>Diagnosi</term>

<trimed:identifier>TPLT18168</trimed:identifier>

<basic:transaction>origination</basic:transaction>

<basic:responsibility>Student TPLT</basic:responsibility>

<date>2018</date>

<note></note>

<min:partOfSpeech>sostantivo</min:partOfSpeech>

<basic:grammaticalGender>femminile</basic:grammaticalGender>

<trimed:grammaticalNumber>invariabile</trimed:grammaticalNumber>

<trimed:pronunciation>dia-gno-si</trimed:pronunciation>

<trimed:etymology></trimed:etymology>

<trimed:variant></trimed:variant>

<trimed:acronym></trimed:acronym>

<trimed:fullForm></trimed:fullForm>

<trimed:abbreviation></trimed:abbreviation>

<trimed:derivative>diagnosticare, diagnostico, diagnosta</trimed:derivative>

```
                <trimed:collocation>fare la diagnosi, formulare una diagnosi</trimed:collocation>
                <trimed:phraseologicalUnit></trimed:phraseologicalUnit>

                <trimed:synonym>anamnesi, prognosi, analisi, accertamento</trimed:synonym>
                <trimed:hypernym></trimed:hypernym>
                <trimed:hyponym></trimed:hyponym>
                <basic:crossReference></basic:crossReference>

                <basic:context></basic:context>
                <basic:externalCrossReference>https://www.airc.it/cancro/informazioni-tumori/
                guida-ai-tumori/tumori-neuroendocrini</basic:externalCrossReference>
                <basic:source></basic:source>

                <trimed:register>specializzato</trimed:register>
                <trimed:commonName>Diagnosi</trimed:commonName>
                <trimed:scientificName>Diagnosi</trimed:scientificName>
                <trimed:mesh>https://meshb.nlm.nih.gov/record/ui?ui=D003933</trimed:mesh>
                <trimed:snomed>https://browser.ihtsdotools.org/?
                perspective=full&amp;conceptId1=439401001&amp;edition=MAIN/
                2020-03-09&amp;release=&amp;languages=en</trimed:snomed>
                <trimed:icpc2Code></trimed:icpc2Code>
                <trimed:icd10Code></trimed:icd10Code>

            </termSec>
          </langSec>

        </conceptEntry>
      </body>
    </text>
</tbx>
```