Aligning Biomedical Terminologies in French: Towards Semantic Interoperability in Medical Applications

Tayeb Merabti[1], Lina F. Soualmia[1,2], Julien Grosjean[1],
Michel Joubert[3] and Stefan J. Darmoni[1]
[1]CISMeF, Rouen University Hospital, Normandy & TIBS, LITIS EA 4108,
Institute of Biomedical Research, Rouen
[2]LIM & Bio EA 3969, Paris XIII University, Sorbonne Paris Cité, Bobigny
[3]LERTIM EA 3283, Faculty of Medicine, Marseilles
France

1. Introduction

In health, there exist practically as many different terminologies, controlled vocabularies, thesauri and classification systems as there are fields of application. In fact, terminologies play important roles in clinical data capture, annotation, reporting, information integration, indexing and retrieval. These knowledge sources have mostly different formats and purposes. For example, among many other knowledge sources, the Systematized NOmenclature of MEDicine International (SNOMED Int) is used for clinical coding, the French CCAM for procedures, the 10th revision of the International Classification of Diseases (ICD10) and the Anatomical Therapeutic Chemical (ATC) Classification for drugs are used for epidemiological and medico-economic purposes and the Medical Subject Headings (MeSH) thesaurus for indexing bibliographic databases. Given the great number of terminologies, existing tools, such as search engines, coding systems or decision support systems, are limited in dealing with "syntactic" and "semantic" divergences in spite of their great storage capacity and quick processing of data. Faced with this reality and the increasing need to allow cooperation with/between the various health actors and their related health information systems, it appears necessary to link and connect these terminologies to make them "interoperable". The objective is to allow the different actors to speak the same language while using different representations of the same things. As it is essential to render these terminologies "interoperable", this involves establishing a joint semantic repository to allow effective interaction with a minimum loss of meaning. This semantic interoperability requires a shared model, i.e. a common representation of terms and concepts, whatever the original terminology or repository is but it also requires the development of methods to allow connection between equivalent terms or relations from each terminology.

Various studies have investigated the implementation of platforms to achieve interoperability between health terminologies. The Unified Medical Language System (UMLS), developed by the US National Library of Medicine since 1986 (Lindberg et al., 1993), is one such project.

Currently, it is considered as the largest existing metathesaurus. However, the UMLS does not make semantically integrated terminology interoperable but rather provides rich health knowledge sources that can potentially be used towards mapping or connection identification. Other studies were interested in the issue of providing terminology servers in the health domain (Chute et al., 1999; Rector et al., 1997). The use of multiple terminologies is recommended to increase the number of lexical and graphical forms of a biomedical term recognized by a search engine. For this reason, in France, since 2005, the Catalog and Index of Health Resources in French (CISMeF) has evolved from a mono-terminology approach using MeSH main headings and subheadings to a multiple terminologies paradigm using, in addition to the MeSH thesaurus, several vocabularies and classifications that deal with various aspects of health.The overall CISMeF Information System (CISMeF_IS), includes multiple-terminologies indexing (Pereira et al., 2008), multi-terminology information retrieval (Sakji et al., 2009; Soualmia et al., 2011) and integrates of several terminologies (n=32) in the CISMeF terminology database. The CISMeF team has created a Health Multi-Terminology Portal (HMTP) largely inspired by the most recent advances in semantic web technologies (Darmoni et al., 2009a). Besides platforms, terminology servers and other computer systems for semantic interoperability, there are significant challenges in developing automated and semi-automated approaches for identifying direct and indirect relations between terms i.e. alignments. Aligning different terminologies by determining relations is a hard task regardless of the research field, whether in Information Science (Zeng & Chan, 2004), matching database schemas (Doan et al., 2004) or aligning ontologies (Euzenat & Shvaiko, 2007). In addition to heterogeneity formats, two problems complicate the alignments between terminologies. Firstly, the informal processing of relations in the terminology which makes several definitions ambiguous (Sarker et al., 2003). Unfortunately, this problem remains difficult to solve because it requires changes in the logical construction of each original terminology: hierarchical relationships, synonymy relations or related relations. The second problem consists in making these approaches automatic. In fact, most of the existing approaches to link terminologies are manual and very time consuming. For example, the manual mapping between ATC and the MeSH thesaurus took more than 6 men.months. Obviously, it is not possible for a team such as CISMeF (n=20) or another team of the same scale to manually produce at least 190 mappings between 32 terminologies $\frac{N(N-1)}{2}$. In this chapter, we aim primarily to contribute to the second problem related to the automation of mapping approaches to identify relations between terminologies. The remainder of the chapter is organized as follows: in section 2 we start by a panel of several biomedical terminologies (including classifications, controlled vocabularies, taxonomies, . . .etc).

Some projects (UMLS and the HMTP) for integrating medical terminologies and ontologies are described in the section 3. The section 4 is devoted to background on terminology and on ontology alignments methods, mainly semantic and syntactic ones. The methods we propose are developed in the section 5. Alignments of specific terminologies are presented in the section 6 and the section 7 displays the global results we have obtained. The section 8 gives several uses of the alignments through the HMTP, mainly for information retrieval and automatic translation. Finally some related work and discuss the results we have obtained and conclude this study in sections 9 and 10.

2. Panel of biomedical terminologies and their use

2.1 Terminology definition

A terminological system links together concepts of a domain and gives their associated terms, and sometimes their definition and code. It might take the designation of *terminology*, *thesausrus*, *controlled vocabulary*, *nomenclature*, *classification, taxonomy* or *ontology*. In (Roche, 2005), terminology was defined as a set of words. A more precise definition of terminology was given in (Lefevre, 2000): "Terminologies are a list of terms of one area or a topic representing concepts or notions most frequently used or most characteristic". Thereby, the content and the structure of a terminology depend on the function for which this terminology will be used.

A terminology in which the terms are for example organized alphabetically and in which the concepts may be designed with one or several synonyms is a *thesaurus*. When the terms are associated to definitions, it constitutes a *controlled vocabulary*. A *nomenclature* is a terminology in which the terms are composed according to pre-existing rules. When hierarchical relations are introduced between concepts, it is a classification. A *classification* is the exhaustive organization of the concepts of a domain into classes, according to their distinctive characteristics. The classes are mutually exclusive and organized hierarchically from the most generic to the most specific. In classifications, one can find classes denoted "No Otherwise Specified" which gather terms that cannot be classified elsewhere. A *taxonomy* is a classification in which the classes have only hierarchical relations of generic type.

In medical terminologies, specific terms are used to specify concepts of the domain. Relations can also exist between terms. For example, generalization and specialization relations (is-a) exist in several terminologies to rank terms from the more general to the more specific, and partitive ones (part-of) designs which term designates a part-of another one. In terminologies, concepts can be designated by several different terms. A Preferred Term (PT) is the term describing a unique medical concept in terminology. The PT is defined as less ambiguous, more specific and self-descriptive as possible. As a continuum with terminology, an ontology is a "formal, explicit specification of a shared conceptualization for a domain of interest" (Gruber, 1993). Usually, an ontology is organized by concepts and identifies all possible inter-relations. Ontologies are used to facilitate communication among domain experts and between domain experts and knowledge-based systems. This is done to reflect the expert view of a specific domain. The difference with terminology, is mainly in knowledge representation language, which is formal in the case of ontology.

2.2 The main medical terminologies

In this section we describe several terminologies. As explained in the introduction, each terminology is developed for a particular use. The following terminologies are the most known in the domain of health:

- the main thesaurus used for medical information is the Medical Subject Headings (MeSH®) (Nelson et al., 2001) maintained by the U.S. National Library of Medicine. It consists of a controlled vocabulary used for indexing the content of health documents and it is available in 41 languages, including 26,000 MeSH Descriptors, 83 MeSH Qualifiers and 200,000 MeSH Supplementary Concepts (MeSH SC) ;

- The Systematized Nomenclature Of MEDicine (SNOMED) International is used essentially to describe electronic health records (Côté et al., 1993), and is a standard for electronic health records (Cornet & de Keizer, 2008);
- Medical Dictionary for Regulatory Activities (MedDRA), for adverse effects (Brown et al., 1999);
- Logical Observation Identifiers Names and Codes (LOINC) (Cormont et al., 2011);
- International Union of Pure and Applied Chemistry (IUPAC) for chemical sciences[1];
- Various codes used for drugs and chemical compounds: CAS for chemistry, Brand Names and International Non-proprietary Names (INN) for drugs.

Several terminologies are developed and maintained by the World Health Organization (WHO):

- The International Classification of Diseases, 10th revision (ICD10)[2];
- The Adverse Reactions Terminology (WHO-ART), for adverse effects[3];
- The Anatomical Therapeutic Chemical Classification System (WHO-ATC)[4] for drugs;
- The International Classification for Patient Safety (WHO-ICPS)[5];
- International Classification of Functioning, Disability and Health(WHO-ICF)[6] for handicap.

Concerning diseases, the ORPHANET thesaurus is available in five languages (English, French, Spanish, Italian and Portuguese). It describes rare diseases, including related genes and symptoms (Aymé et al., 1998). The MEDLINEPlus thesaurus (Miller et al., 2000) is a thesaurus for lay people. More formal representations exist. For example :

- Foundational Model of Anatomy (FMA) (Noy et al., 2004; Rosse & Mejino, 2003) which describes anatomical entities.
- Human Phenotype Ontology (HPO) (Robinson & Mundlos, 2010)

In France, the Joint Classification of Medical Procedures (CCAM) (Rodrigues et al., 1997) and ICD10 are mandatory for epidemiological and medico-economic purposes for all private and public health care institutions. The International Classification of Primary Care, Second edition (ICPC2) [7] and French dictionary for outpatients (DRC)[8] are two classifications for family medicine and primary care respectively designed by the World Organization of National Colleges (WONCA), Academies, and Academic Associations of General Practitioners/Family Physicians) and the French Society of Family Medicine (SFMG)). Two French terminologies exist to describe medical devices: LPP[9] and CLADIMED[10]. LPP is

---

[1] http://www.iupac.org/
[2] http://www.who.int/classifications/icd/en/
[3] http://www.umc-products.com
[4] http://www.whocc.no/atcddd/
[5] http://www.who.int/patientsafety/implementation/taxonomy/
[6] http://www.who.int/classifications/icf/en/
[7] http://www.who.int/classifications/icd/adaptations/icpc2/
[8] http://www.sfmg.org/outils_sfmg/dictionnaire_des_resultats_de_consultation-drc/
[9] http://www.codage.ext.cnamts.fr/codif/tips/index_presntation.php?p_site=AMELI
[10] http://www.cladimed.com

the list of medical devices from the French National Health Insurance and CLADIMED is a five level classification for medical devices, based on the ATC classification approach (same families). Devices are classified according to their main use and validated indications.

Another original way to represent medical concepts is the use of a graphical language based on pictograms, icons and colors with compositional rules (Lamy et al., 2008). We have presented a few examples of existing terminologies and their use. Development of techniques to allow semantic interoperability between these knowledge sources of heterogeneous formats and contents. In the following section we describe projects developed in the US and in France that have proposed efficient ways to connect several terminologies of different use, languages and formats.

3. Integrating medical terminologies for interoperability

3.1 The Unified Medical Language System (UMLS) Project

The richest source of biomedical terminologies, thesauri, classifications is constituted by the Unified Medical Language System (UMLS) Metathesaurus (Lindberg et al., 1993) initiated by the U.S. NLM (National Library of Medicine) with the purpose of integrating information from a variety of sources. It is a way of overcoming two major barriers to efficient retrieval of machine-readable information: (i) the different expression of the same concepts in different machine-readable sources and by different people; (ii) the distribution of useful information between databases and systems. The purpose of UMLS is to facilitate the development of computer systems that use the biomedical knowledge to understand biomedicine and health data and information. To that end, the NLM distributes two types of resources for use by system developers and computing researchers:

- The UMLS Knowledge Sources (databases) integrates over 2 million names for some 900,000 concepts from over 154 biomedical vocabularies from 60 families of vocabularies, as well as 12 million relations between these concepts used in patient records, administrative data, full-text databases and expert systems (Bodenreider, 2004). There are three UMLS Knowledge Sources: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon.
- Associated software tools to assist developers in customizing or using the UMLS Knowledge Sources for particular purposes. Some of the tools included are for example MetamorphoSys (a tool for customization of the Metathesaurus), Lexical Variant Generator (LVG) for generation of lexical variants of concept names or MetaMap (for extraction of UMLS concepts from texts).

The UMLS Metathesaurus is a very large, multi-purpose, and multilingual vocabulary database that contains information about biomedical and health-related concepts, their various names, and the relationships between them. It is built from the electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, biomedical literature indexing and cataloging, and health services research. All the terminologies are under a common representation. The Metathesaurus creates concepts from the various sources and assigns each concept a Concept Unique Identifier (CUI). A CUI may refer to multiple terms from the individual terminologies. These concepts are labeled with Atomic Unique Identifiers (AUIs). For example, the AUI *Cold Temperature* [A15588749] from MeSH and the AUI *Low Temperature*

[A3292554] from SNOMEDCT are mapped to the CUI *Cold Temperature* [C0009264]. Ambiguity arises in the Metathesaurus when a term maps to more than one CUI. For example, the term *cold* maps to the CUIs *Cold Temperature* [C0009264], the *Common Cold* [C0009443], *Cold Sensation* [C0234192], *Chronic Obstructive Lung Disease* [C0024117], or *Colds homeopathic medication* [C1949981] the meaning of which is correct depending on the context in which the term is used. Concept Unique Identifiers CUIs in the Metathesaurus denote possible meanings that a term may have in the Metathesaurus. A CUI is expressed by specific attributes that define it such as its: preferred term, associated terms (synonyms), concept definition, related concepts. For example, the CUI C0009264 has the preferred term *Cold Temperature*. The definition of *Cold Temperature* [C0009264] is: Having less heat energy than the object against which it is compared; the absence of heat. Some of the terms associated with *Cold Temperature* [C0009264] are: *Cold Temperature, Low Temperature, Cold Thermal Agent* and *Cold*. There are two different types of relations that can exist between concepts, subsumption relations (is-a) such as parent/child, and other relations such as siblings. For example, the parent of *Cold Temperature* [C0009264] is *Temperature* [C0039476] and one of its siblings is *Hot Temperature* [C2350229].

Among the 154 biomedical vocabularies, the UMLS Metathesaurus includes only six (6) French terminologies: the MeSH, ICD10, SNOMED Int, WHO-ART, ICPC2 and MedDRA. Nevertheless, only four (4) terminologies are included with their French version in UMLS Metathesaurus (MeSH, WHO-ART, WHO-ICPC2 and MedDRA). However, several translations have already been added, such as MEDLINEPlus (Deléger et al., 2010) and partially LOINC and FMA (Merabti et al., 2011). The SPECIALIST Lexicon provides the lexical information of many biomedical terms. The information available for each word or term records includes syntactic, morphological and orthographic information. This lexical information is very useful for natural language processing systems, specifically for the SPECIALIST NLP (Natural Language Processing) System. However, the SPECIALIST Lexicon contains only English biomedical terms and general English terms and the associated NLP tools stands for English. The Semantic Network provides a categorization of Metathesaurus concepts into semantic types and relationships between semantic types. It provides a set of useful relationships between concepts represented in the Metathesaurus and a consistent categorization of all these concepts The current release of the Semantic Network contains 135 semantic types and 54 relationsh. A semantic type is a cluster of concepts that are meaningfully related in some way. For example, the semantic type of *Cold Temperature* is *Natural Phenomenon or Process*, whereas *Temperature* is assigned the semantic type *Quantitative Concept*. A concept may be assigned more than one semantic type. Nonetheless, the Metathesaurus does not allow interoperability between terminologies since it integrates the various terminologies as they stand without making any connection between the terms in the terminologies other than by linking equivalent terms to a single identifier in the Metathesaurus. For example the concept *Addison's disease* [C0001403] corresponds to :

3.2 Health Multi Terminology Portal (HMTP)

Since 2005, the Catalog and Index of Health Resources in French (CISMeF) evolved from a mono-terminology approach using the MeSH main headings and subheadings to a multiple terminologies paradigm using, in addition to the MeSH thesaurus, several vocabularies and classifications that deal with various aspects of health. The CISMeF team has created a Health Multi-Terminology Portal (HMTP) largely inspired by the most recent advances in

Addison's disease	[SNOMEDCT]	PT363732003
Addison's Disease	[MedlinePlus]	PT1233
Addison Disease	[MeSH]	D000224
Addison's disease	[SNOMED CT]	PT363732003
Addison's Disease	[MedlinePlus]	T1233
Addison Disease	[MeSH]	D000224
Bronzed disease	[SNOMED Int 1998]	DB-70620
Deficiency; corticorenal, primary	[ICPC2-ICD10 Thesaurus]	THU021575
Primary Adrenal Insufficiency	[MeSH]	D000224
Primary hypoadreanlism syndrome, Addison	[MedDRA]	10036696

web technologies (Darmoni et al., 2009b; Grosjean et al., 2011). The HMTP includes all the terminologies listed in section 2.2 and others related to drugs : the International Union of Pure and Applied Chemistry (IUPAC) for chemical sciences, various codes used for drugs and chemical compounds: CAS for chemistry, Brand Names and International Non-proprietary Names (INN) for drugs, CIS, UCD, and CIP for French drugs.

The HMTP includes also a CISMeF thesaurus (Douyère et al., 2004), which is an extension to the MeSH thesaurus, includes 130 metaterms (super-concepts to unify MeSH terms of the same medical discipline), 300 resource types (adaptation to the Internet of the publication types), over 200 predefined queries and the translation of 12,000 MeSH Scope Notes (8,000 manually and the rest semi-automatically). To fit all the terminologies into one global structure and allow semantic interoperability, a generic model compliant with the terminology ISO model was designed. It was established around the "Descriptor" which is the central concept of the terminologies (aka "keyword"). The HMTP is a "Terminological Portal" connected to generic model database to search terms among all the health terminologies available in French (or in English and translated into French) and to search it dynamically. The ultimate goal is to use this search via the HMTP in order to: (i) manually or automatically index resources in the CISMeF catalog; (ii) allow multi-terminology information retrieval (Darmoni et al., 2009b; Soualmia et al., 2011).

It can also be very useful in teaching or performing audits in terminology management. Currently, the HMTP allows users to access 32 terminologies and classifications. Some of those are included in the UMLS meta-thesaurus (n=9) but the majority are not (n=23) such as the ORPHANET thesaurus (Aymé et al., 1998), DRC (Ferru & Kandel, 2003), IUPAC[11]. Table 1 lists most of the terminologies included in the HMTP and table 2 displays the number of descriptors and relationships included.

4. Semantic integration through alignments

4.1 Methods for aligning terminologies and ontologies

Ontology alignment is the task of determining correspondences between concepts of different ontologies. A set of correspondences is also called an alignment (Euzenat & Shvaiko, 2007).

---

[11] IUPAC: http://www.iupac.org

Terminology	HMTP	UMLS
CCAM	Included (Fr and En)	
CISMeF	Included (Fr and En)	
Codes used for drugs	Included (Fr and En)	
DRC	Included (Fr and En)	
FMA	Included (Fr and En)	Included (En)
ICD10	Included (Fr and En)	Included (En)
IDIT	Included (Fr)	
IUPAC	Included (Fr and En)	
LOINC	Included (Partially translated Fr, En)	Included (En)
MedDRA	Included (Fr and En)	Included (Fr and En)
MEDLINEPlus	Included (Fr and En)	Included (En)
MeSH	Included (Fr and En)	Included (Fr and En)
NCCMERP	Included (En)	
ORPHANET	Included (Fr and En)	
PSIP Taxo.	Included (En)	
SNOMED International	Included (Fr and En)	Included (En)
UNIT	Included (Fr and En)	
VCM	Included (Fr)	
WHO-ART	Included (Fr and En)	Included (Fr and En)
WHO-ATC	Included (Fr and En)	
WHO-ICF	Included (Fr and En)	Included (En)
WHO-ICPC2	Included (Fr and En)	Included (Fr and En)
WHO-ICPS	Included (Fr and En)	

Table 1. List of the most represented terminologies included in the HMTP.

Terminologies	32
Terms/Concepts	980,000
Synonyms	2,300,00
Definitions	222,800
Relations and hierarchies	400,000

Table 2. Main figures of the Health Multi-Terminology Portal (November 2011).

Historically, the need for ontology alignment arose out of the need to integrate heterogeneous databases developed independently and thus each having their own data vocabulary. As terminology is a kind of ontology the definition of Euzenat stands for *Terminology Alignment*: the task of determining correspondences, i.e. alignments, between terms. Various studies have investigated automatic and semi-automatic methods and tools to map between medical terminologies to make them "interoperable" . The terminologies themselves are unaffected by the alignment process. Alignment techniques are of particular importance because the manual creation of correspondences between concepts or between terms is excessively time consuming. According to (Shvaiko & Euzenat, 2005) there are two major dimensions for similarity: the syntactic dimension and the semantic dimension. Syntactic dimension is based

on lexical methods and semantic dimension is based on structural and semantic properties of terminologies (Euzenat & Shvaiko, 2007).

4.1.1 Lexical methods

Lexical methods are based on the lexical properties of terms. These methods are straightforward and represent a trivial approach to identifying correspondences between terms. The use of such methods in the medical domain to achieve mappings was motivated by the fact that most terminologies share many similar terms.

String-based Methods In these methods, terms or (labels) are considered as sequences of characters. A string distance is determined to compute a similarity degree. Some of these methods can skip the order of characters. Examples of such distances, also used in the context of information retrieval, are: the Hamming distance (Hamming, 1950), the Jaccard distance (Jaccard, 1901), Dice Distance (Salton & McGill, 1983). On the other hand, a family of appropriate measures known as "Edit distance", takes into account the order of characters. Intuitively, an edit distance between two strings is defined as being the minimum number of character inserts, deletes and changes needed to convert one string to another. Levenshtein distance (Levenshtein, 1966) is one example of such distances. It is the edit distance with all costs equal to 1. Another example of such distance is the SMOA distance (Stoilos et al., 2005) which is based on the idea that the similarity between two strings depends on their commonalities and differences. However, these methods can only quantify the similarity between terms or labels. Thus, they produce low (or no) similarity between synonyms term with different structures. For example, the two words "pain" and "Ache" are synonyms, *i.e.* related semantically as being the same thing, but all the distances presented above cannot identify any links between these two terms. Conversely, these methods find significant similarity between different terms (false positive), such as: "Vitamin A" and "Vitamin B".

Language-based Methods In these methods, terms are considered as words in a particular language. They rely on NLP tools to help the extraction of the meaningful terms from a text. These tools exploit morphological properties of words. We distinguish methods which are based on normalization process from those which exploit external knowledge resources such as dictionaries.

  Normalization methods Each word is normalized to a standardized form that can be easily recognized. Several linguistic software tools are developed to quickly obtain a normal form of strings : (i) tokenization consists in segmenting strings into sequences of tokens by eliminating punctuation, cases, blank characters; (ii) the stemming process consists in analyzing the tokens derived in the tokenization process to reduce them to a canonical form; (iii) the stop words elimination consists in removing all the frequent short words that do not affect the sentences or the labels of terms, phrases such as "a", "Nos", "of"...etc

  External-based methods These methods use external resources, such as dictionaries and lexicons. Several linguistic resources exists to found possible mappings between terminologies exist. These methods form the basis of the lexical tools used by the UMLSKS API (section 3.1). They were combined with synonyms from other external resources to optimize mapping to the UMLS. Another external resource largely used in the biomedical field is the lexical database WordNet (Fellbaum, 1998).

4.1.2 Semantic (or structural) methods

These methods use the structural properties of each terminology to identify possible correspondences between terms. They consider terminologies as graph were nodes represent terms and edge represent relations established in the terminology between these terms. Most medical terminologies can be represented as graph. Furthermore, these techniques can also be combined with lexšical techniques. The work presented in (Bodenreider et al., 1998) is a good example that illustrating the use of terminology relations to map terms not mapped with lexical methods. This algorithm used the semantic relationships between concepts from different terminologies included in the UMLS. In parallel with the structural properties of each terminology, semantic methods used also semantic similarities to find the closest term. The main technique consists in computing the number of edges between terms to determine a distance between them. The famous similarity distance is the Wu-Palmer distance (Wu & Palmer, 1994). This similarity is defined according to the distance between two terms in the hierarchy and also by their positions from the root. Unlike these traditional edge-counting approaches, other methods calculate the similarity according to the most information that two terms share in a hierarchical structure such as: Lin similarity (Lin, 1998) for example, this similarity was combined with a statistical similarity used to compute semantic similarity between CISMeF resources (Merabti et al., 2008). These similarities can be used to find possible connections between terms or concepts from different hierarchical terminologies, such MeSH or SNOMED Int for example.

4.2 Methods for evaluation of mapping results

Although fully automatic alignment might appear as the solution of choice for the interoperability of semantic systems, results provided by fully automatic methods are rarely of sufficient quality. In parallel to mapping methods, several techniques and methods were proposed to evaluate the mapping results produced by several systems. As defined in (Euzenat & Shvaiko, 2007), the goal of evaluation is to improve the mapping method and to give the user the best tool and method possible for the task. The main evaluation methods are based on the appropriateness and quality of the results, using a Likert scale or measures such as precision, recall, the F-measure and the of mapping. In (Ehrig & Euzenat, 2005) the authors proposed a framework for generalizing precision and recall and in (Euzenat, 2007) the author proposed a semantic precision and recall. These improvements were analyzed in (David & Euzenat, 2008) where more adaptations of these two measures to normalized mapping are proposed. In (Euzenat et al., 2011) one can find a panel of systems and results concerning the Ontology Alignment Evaluation Initiative. As in Information Retrieval systems evaluation, this type of evaluation needs a gold standard (GS) dataset. The problem is that these datasets are not available or easy to find or build as stated in (Euzenat et al., 2011). This is why the majority of evaluations used for our studies described hereafter are based on Likert scales where an expert manually evaluates a small set of mapping results according to specific levels. Nevertheless, the necessity of involving humans in the alignment process using visual interfaces has been outlined in (Kotis & Lanzenberger, 2008) within a discourse on ontology alignment challenges. On the same issue, as argued in (Granitzer et al., 2010) visual interfaces can address efficiently the problem of evaluating automatic alignment systems to take advantage of human cognitive capabilities and provide intuitive overview, navigation and detail analysis. Therefore, from next year we are going to offer to experts an evaluation

tool connected to our databases to facilitate the evaluation of each automatic mapping. We think that regulated use of this tool can allow us to build a large dataset with valid and non valid mappings between terminologies that can be used to improve our methods.

5. Proposed methods for aligning medical terminologies

In this section we detail the methods we have developed for aligning terminologies included in the UMLS and HMTP described in section 3. We also detail the methods we have applied to evaluate the mapping results. Two automatic mapping approaches are implemented in the HMTP: conceptual and lexical approach. The former uses the UMLS Metathesaurus to map the terminologies included in the UMLS, whereas the latter exploits natural language processing tools to map terminologies whether or not they are included in the UMLS.

5.1 Conceptual approach

This approach is possible if each term to be mapped is included in the Metathesaurus (Joubert et al., 2009). The principle of the method is based on the conceptual construction of the UMLS Metathesaurus. Three types of mapping could be derived: "ExactMapping", "BroaderMapping" and/or "NarrowMapping" and "CloseMapping" (see Table 3 for examples). This method is inspired by the SKOS (Simple Knowledge Organization System) definitions of mapping properties[12]. Let t_1 and t_2 two terms belong to T_1 and T_2, two terminologies respectively. Suppose CUI_1 and CUI_2, the respective projections of t_1 and t_2 in the Metathesaurus, then t_1 and t_2 could be aligned if:

- $CUI_1 = CUI_2$, this corresponds to the "Exact Mapping".
- there is a parent of t_1 or t_2 which maps t_2 or t_1 respectively, this corresponds to "Broad Mapping" and/or "Narrow Mapping": these are used to state mapping links through hierarchies.
- there is explicit mapping between CUI_1 and CUI_2, this corresponds to the non-transitive "Close Mapping": two concepts are sufficiently similar that they can be used interchangeably.

The algorithm is carried out sequentially and stops when a candidate term for mapping is found. As an application of this, even if an explicit mapping comes from other terminologies in UMLS, e.g. ICD-9-CM and SNOMED CT (Imel, 2002) not part of terminologies under consideration, explicit mappings between two terminologies can be "reused" for other terminologies by using the UMLS concept structure (Fung & Bodenreider, 2005).

5.2 Lexical approach

In this approach, Natural Language Processing (NLP) tools adapted for the English and French languages) are used to link terms from different terminologies in the HMTP. The lexical approach allows us to find a term in the target terminology that is the most lexically similar to a given term in a source terminology.

---

[12] World Wide Web Consortium Simple Knowledge Organization System: `www.w3.org/2004/02/skos`

Type of relation	Source term (Terminology)	Target Term (Terminology)
Exact Mapping	Congenital bladder anomaly (MedDRA)	Congenital anomaly of the bladder, nos (SNMI)
Close Mapping	Diseases of lips (ICD10)	Ulcer of lip (SNMI)
BT-NT Mapping	Hepatic insufficiency (MeSH)	Liver disease, nos (ICPC2)

Table 3. Examples for each type of conceptual mapping.

5.2.1 Lexical approach for medical terminologies in French

This approach uses a French NLP tool and mapping algorithms developed by the CISMeF team to map French medical terminologies (Merabti, 2010; Merabti et al., 2010a;b). These tools were initially developed in previous works for information retrieval (Soualmia, 2004) and extended to link terms in multiple French medical terminologies:

- Remove stop words: frequent short words that do not affect the phrases such as "a", "Nos", "of", etc are removed from all terms in all terminologies in the HMTP.
- Stemming, a French stemmer provided by the "Lucene" software library which proved to be the most effective for automatic indexing using several health terminologies (Pereira, 2007). Mapping used by this approach may provide three types of alignments between all terms:
- Exact correspondence: if all words composing the two terms are exactly the same.
- Single to multiple correspondences: when the source term cannot be mapped by one exactly target term, but can be expressed by a combination of two or more terms.
- Partial correspondence: in this type of mapping only a part of the source term will be mapped to one or more target terms.

Examples for each type of mapping are given in Table 4. In this work, we describe only exact correspondences.

Type of correspondance	Source term (Terminology)	Target Term(s)(Terminology)
Exact	Syndrome de Marfan "Marfan Syndrome"(MeSH)	Syndrome de Marfan "Marfan's Syndrome" (MedDRA)
Single to Multiple	Albinisme surdité "Albinism- deafness syndrome" (ORPHANET)	Albinisme "Albinism" (MeSH) and (+) Surdité "Deafness" (SNMI)
Partial	Chromosome 14 en anneau "Ring chromosome 14" (ORPHANET)	Chromosome humain 14 "Chromosome 14" (MeSH)

Table 4. Examples of the three types of mappings using the French lexical approach.

5.2.2 Lexical approach for medical terminologies in English

In this approach we use lexical tools in English developed by the NLM (Browne et al., 2003) and included in the Lexical tool of the UMLS (see section 3.1). These tools were designed to aid users in analyzing and indexing natural language texts in the medical field (McCray et al., 1994; Peters et al., 2010). They include essentially :

- the LVG (Lexical Variant Generator): a Multi-function tool for lexical variation processing;
- Norm[13]: a program used to normalize English terminologies included in the UMLS ;
- WordInd: a tool used to tokenize terms into words.

In this work we have used the normalization program ("Norm"). The normalization process involves stripping genitive marks, transforming plural forms into singular, replacing punctuation, removing stop words, lower-casing each word, breaking a string into its constituent words, and sorting the words into alphabetic order. We have considered here only the exact correspondences. This type of mapping is easy to evaluate in English and the "not exact" correspondence will be useful for the translation of English terms into French. Several tools based on these techniques were used to map between medical terminologies. As an example, the authors in (Wang et al., 2008) used tokenisation and stemming techniques to map ICPS-2 with the SNOMED CT. It is also the case for the lexical techniques proposed by the NLM in the UMLSKS API. The NLM also created (Aronson, 2001) a tool to identify biomedical concepts from free textual input and map them into concepts from the UMLS. Authors in (Johnson et al., 2006) used the Lucene API to found relations between Gene Ontology (GO) and three other biomedical ontologies.

5.3 Structural approach

This approach is based on hierarchical relations and was used to align the remaining terms not mapped by the lexical approach. This mapping provides two types of correspondences:

- BroadMapping: when the remaining term has at least one parent (hierarchical relation broader than) mapped to at least one term.
- NarrowMapping: when the remaining term has at least one child (hierarchical relation narrower than) mapped to at least one term.

The work presented in (Bodenreider & McCray, 1998) is a good example that illustrates the use of the terminology relations to map terms not mapped with the lexical methods. This algorithm exploit the semantic relationships between concepts from different terminologies included in the UMLS.

6. Cases studies

In this section we present some cases of alignments between medical terminologies, essentially in French, to the UMLS, HMTP or other terminologies.

---

[13] National Library of Medicine: Lexical Tools:
  `http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/2010/docs/userDoc/index.html`

6.1 Aligning the ORPHANET thesaurus to the MeSH thesaurus

In order to align the ORPHANET thesaurus, which describes rare diseases to the MeSH thesaurus, we have compared two methods. The first one uses the UMLS and an external manual alignment of ORPHANET terms to ICD10 codes. The second one uses only lexical-based approach without using the UMLS to make a direct and an automatic alignment between ORPHANET and MeSH. We also provide an evaluation and a comparison of these two methods. The MeSH thesaurus was chosen as the target terminology for comparing alignment strategies for two main reasons:

- the ORPHANET team needs to map each ORPHANET term to a MeSH term to allow a contextual link between an ORPHANET Web page for one ORPHANET rare disease (e.g. Marfan syndrome) and one corresponding PubMed query. The CISMeF team has strong experience with the MeSH thesaurus. Therefore, the evaluation will conducted done by a CISMeF expert.
- the MeSH thesaurus is the second largest terminology available represented in the UMLS and it freely available in the HMTP. Nevertheless, ORPHANET is now aligned to all French and English terminologies available in the HMTP and several relations from this terminology are also available (not freely) in the HMTP.

6.1.1 Methods

The first method "Manual ORPHANET-ICD10 link-based alignment" is based on the external manual alignment between ORPHANET and ICD10 terms performed by the ORPHANET team. In this approach, the link provided by the UMLS Metathesaurus between ICD10 and MeSH is used. Hence, an effective alignment exists between two terms ICD10 and MeSH if these terms share the same UMLS Concept Unique Identifier (CUI) in the Metathesaurus. For example, there is an effective alignment between the ICD10 term "Cushing syndrome" (Code: E24) and the MeSH term "Cushing syndrome" since they share the same UMLS Concept CUI: C0010481)(Table 5).

ORPHANET Term	ICD10 term	MeSH term
Cushing syndrome	SCushing's syndrome	Cushing Syndrome
Ichthyosis, X-linked	X-linked ichthyosis	Ichthyoses, X-Linked
Muscular dystrophy, Duchenne and Becker types	Muscular dystrophy	Muscular Dystophies

Table 5. Example of UMLS and manual ORPHANET-ICD10 links based mapping.

The second method is the "Lexical-based alignment" which is described in the section 5.2 This method allows us to find a term in the target terminology (MeSH) that is the most lexically similar, from a given term in the source terminology (ORPHANET). We have however also used a structural approach to align the remaining ORPHANET terms to the MeSH.

6.1.2 Evaluation & comparison

To evaluate the two methods, four sets of correspondences were derived from the results of the two methods applied to 2,083 ORPHANET terms manually aligned to the ICD10:

- First set: The correspondences obtained by the first strategy "UMLS and manual ORPHANET ICD10 link-based alignment" and not by the second "lexical-based approach" (only manually found)
- Second set: The correspondences found by the second method and not by the first (Only Lexical-based mapping found)
- Third set: The discrepant correspondences found by both methods for the same ORPHANET term. For example, for the ORPHANET term "Tangier disease" the two methods found two different MeSH terms, the MeSH term "Hypolipoprotenemia" with the first method and the MeSH term "Tangier disease" with the second.
- Fourth set: The correspondences found with both methods (the same correspondences).

A sample of 100 correspondences, randomly determined, from each set was evaluated by a physician (SJD), head of the CISMeF team. The following terms were used to describe the quality of each mapping result: (i)"relevant" the mapping between one MeSH term and one ORPHANET term was rated as correct; (ii) "non-relevant" when the mapping between MeSH and ORPHANET terms was considered by the expert as not correct; (iii) "BT-NT" the ORPHANET term was rated as broader than the MeSH corresponding term; (iv) "NT-BT" the ORPHANET term was rated as narrower than the MeSH corresponding term. For example, "Duchenne and Becker muscular dystrophy" is narrower than "muscular dystrophies" and (v) "Sibling" when the MeSH corresponding and ORPHANET term are siblings (from the MeSH point of view). For example, "Cryptophthalmia, isolated" is evaluated as the sibling of "microphthalmos".

6.1.3 Results

For the UMLS and manual ORPHANET-ICD10 link-based alignment: Among the 2,083 ORPHANET terms (28% of all ORPHANET terms) manually aligned to at least one ICD10 code, 619 possible correspondences were found for at least one MeSH term using the UMLS (30% from 2,083). For the lexical-based approach (only limited to the ORPHANET terms manually linked to ICD10), among the 2,083 ORPHANET terms linked manually to at least one ICD10 code, 593 possible correspondences were found for at least one MeSH term (28% from 2,083). However, 1,004 possible correspondences were found to at least one MeSH term (13% from 7,424) when this method was applied to the whole ORPHANET thesaurus. According to the results of each method we obtained:

1. The first set contains 327 correspondences were found only by the "UMLS and manual ORPHANET ICD10 manual alignments" and not by the "lexical-based alignment".
2. The second set contains the 306 correspondences were found only by the "lexical-based alignment".
3. The third set contains the 75 different correspondences were found by both methods with the same ORPHANET term.
4. The fourth set contains the 211 same correspondences were found by both methods.

The results of the evaluation of the correspondences obtained by each strategy independently are displayed in Table 6. Overall 85% of correspondences obtained by method 2 (Lexical-based mapping) are ranked as relevant when only 21% of correspondences are ranked as relevant for the first strategy (UMLS and manual ORPHANET-ICD10 link-based alignment), whereas 32% and 15% of the correspondences obtained by methods 1 and 2 respectively are ranked

as NT-BT (the source term is evaluated as narrower than the target term in the MeSH hierarchy). Table 7 displays the evaluation results for the third set containing different correspondences from the two strategies for the same ORPHANET term. For the first strategy (UMLS and manual Orphanet-ICD10 link-based alignment), overall 39 correspondences are evaluated as "BT-NT" when only 6 correspondences are evaluated as "relevant". For the second method (Lexical-based mapping), overall there are 62 correspondences evaluated as "relevant", whereas 8 correspondences are evaluated as "BT-NT". The results of evaluation for the fourth set containing the same correspondences derived by each method found relevant correspondences in 98% cases and BT-NT relations in 2% cases.

	Relevant	BT-NT	NT-BT	Sibling	Non-relevant
First Set	21	2	32	0	45
Second Set	85	0	15	0	0

Table 6. Evaluation results of the two sets of correspondences (correspondences found by each strategy only).

	Relevant	BT-NT	NT-BT	Sibling	Non-relevant
UMLS and manual ORPHANET-ICD10 link-based mapping	6	39	7	2	21
Lexical-based alignement	62	8	1	1	2

Table 7. Evaluation results of the fourth set of correspondences (for the same ORPHANET term different correspondences).

Using a lexical-based approach (to all HMTP) 4,669 ORPHANET terms were aligned to at least one terminology from the HMTP. From this set of correspondences, 1,433 ORPHANET terms were aligned with at least one MeSH term (30%). On the other hand, from the remaining ORPHANET terms the structural alignments between ORPHANET and all the terms from HMTP provided: 1,513 ORPHANET terms in broader correspondence and 957 ORPHANET terms in NT correspondence. An ORPHANET expert has evaluated the two correspondences: lexical-based and structural. From 100 lexical-based alignments, 99% were evaluated as relevant and from 500 structural alignments 482 were evaluated as relevant, when 16 were evaluated as irrelevant.

6.2 Aligning the CCAM to the UMLS

The objective of this section is to describe an alignment method that may be used to integrate any medical terminology in French in the UMLS Metathesaurus. The alignment method has been used and evaluated to align the CCAM terminology (Classification Commune des Actes Médicaux) for procedures to the UMLS Metathesaurus. The CCAM is a multi-hierarchical structured classification for mainly surgical procedures used in France for reimbursement and policymaking in health care. Each procedure is described by a code using "CCAM Basic

Coding System", which consists of coding: (1) body system/anatomical site or function, (2) action and (3) approach/method (see the Table 8).

NC	Bones of the leg
C	Osteosynthesis
A	Open Approach
NCCA010	Osteosynthesis of tibial diaphysis fracture by external fixing

Table 8. Example of CCAM basic coding.

6.2.1 Method

The alignment method for mapping CCAM codes to UMLS concepts is based on the structure of the CCAM codes. However, it is impossible to assign one or more specific UMLS concept using only CCAM labels. This is mainly due to the length of CCAM labels. Indeed, there are 85% of CCAM labels that are composed of 5 or more than 5 words vs. only 5% of the MeSH descriptors. In this approach, only the first significant three characters that compose the CCAM code according to the anatomic and action axes are aligned with the UMLS Metathesaurus. For example, the CCAM code "MZQH001" that has the label "Arthrography of upper limb with scanography [Arthroscan ofupper limb]", is represented according to the first significant three characters with "Bones, joints and soft tissues of upper limb, multiple locations or not specified + Arthrography". In this context we have used the lexical-based method described in section 5.2 to align the first three characters of each CCAM code. This alignment provides three types of correspondences between all terms in source terminologies and French terms of the UMLS Metathesaurus: (i) exact, (ii) Single to multiple and (iii) partial(see Table 9).

CCAM code	Anatomic axis	Action axis	Corresponding term	Type of mapping
BDHA001	Cornea	biopsy	Biopsy cornea	Exact
AAFA003	Brain	exeresis	Brain and Exeresis	Single to multiple
DGFA013	Aorta	laparotomie	Aorte	Partial correspondance

Table 9. Examples of the three types of mappings using the lexical-based approach.

6.2.2 Evaluation

Evaluation was performed on all correspondences from the "exact" set and for only 100 correspondences from the "Single to multiple" set. We chose only 100 mappings because in most cases the same codes with the same first three characters are mapped to the same terms (HLHH003, HLHH004...). Qualitative evaluation was performed by a physician, expert in CCAM codes and in UMLS. The following terms were used to rate the quality of each

correspondence: (i) "equivalent" the UMLS concept corresponds exactly to the CCAM code; (ii) "BT-NT" when the CCAM code was rated as broader than the UMLS concept according to the label of the CCAM and the preferred terms (PTs) in the UMLS concepts; (iii) "NT-BT" the CCAM code was rated as narrower than the PTs in the UMLS concept; (iv) "incomplete" when the UMLS concept only reflects some part of the CCAM label and (v) "irrelevant" when the correspondence was considered by the expert as incorrect. For example, the correspondence between the CCAM code "HLFA001" (label: "Right hepatectomy, by laparotomy") and the UMLS concept C0193399 (preferred term: "Lobectomy of liver") was rated as NT-BT because the UMLS concept is narrow and less precise than the CCAM label. However, for the "Single to multiple" set, the expert performed the evaluation in two steps: (1) each pair (CCAM axe, UMLS concept) is evaluated independently and (2) the correspondence between the CCAM code and the combination of the UMLS concepts is evaluated in this second phase. For example, evaluating the correspondence between the CCAM code "AAFA003" and the two UMLS concepts: C006104 (preferred term: "Brain") and C0919588 ((preferred term: "Exeresis"), (i) first, the expert evaluates each axe with corresponding UMLS ((Brain, C006104) =equivalent and (Exeresis, C091958) =equivalent)); (ii) second, the expert evaluates the correspondence between the label and the combination of the two UMLS concepts (AAFA003, (C006104, C091958) =NT-BT).

6.2.3 Results

Using this method, there are 5,212 (65%) CCAM codes out of the 7,926 CCAM codes used in this study that provide possible correspondences from the CCAM to French terms in the UMLS. The results of each type of correspondence are displayed in Table 10. There are 2,210 (27.5%) correspondences according to both the anatomic and action axes. In the other hand, there are 1,716 (21%) correspondences according to the anatomic axis alone and 1,286 (16%) correspondences according to the action axis. Overall, 65% of the correspondences "anatomic terms" in the CCAM codes are aligned to at least one UMLS Concept and 37% of the correspondences "action terms" in the CCAM codes are aligned to at least one UMLS Concept. For the set of exact correspondences (n=200), 182 (91%) correspondences between CCAM codes and UMLS concepts were rated as NT-BT and only in 9 cases where they rated as equivalent (see Table 11). For the set of single to multiple correspondences (n=100), 61 (61%) and 44 (44%) of the anatomic and the action axes respectively are equivalent to at least one UMLS concept. According to this type of correspondence, 27 (27%) correspondences between CCAM code and at least one UMLS concept were rated as exactly equivalent, when 54 were rated as NT-BT (see Table 12).

Type of mapping	Number of mappings
Exact	200(2.5%)
Single to multiple	2,010(25%)
"Exact" Partial mapping	3,002(37.8%)

Table 10. Results of each correspondence type.

Relevant	BT-NT	NT-BT	Incomplete	Irrelevant	Total
9 (4.5%)	0 (0%)	182 (91%)	3 (1.5%)	6 (3%)	200

Table 11. Evaluation results of the "exact" correspondence set.

Single to multiple mapping	Equivalent	BT-NT	NT-BT	Incomplete	irrelevant
Anatomic	61 (61%)	1 (1%)	29 (29%)	9 (9%)	0 (0%)
Action	44 (44%)	0 (0%)	49 (49%)	1(1%)	6 (6%)
Combination	27 (27%)	0 (0%)	54 (54%)	10 (10%)	9 (9%)

Table 12. Evaluation results of the "Single to Multiple" correspondence set (n=100).

7. Global results

7.1 Conceptual approach

There are 199,786 correspondences exist between at least two French terms from UMLS (25,833 (ExactMapping), 69,085 (CloseMapping) and 104,868 (Broader and /or NarrowerMapping)). In contrast, from the 25,833 terms rated "Exactly", 15,831 come from SNOMED International whereas only 296 come from ICPC2 (Table 13). The three types of correspondences ("Exact", "Broader" and/or "Narrow" and "Close") are included in the HMTP (see Figure 1).

Terminology	Number of terms mapped
ICD10	3,282 (35%)
ICPC2	296 (39%)
MedDRA	5,700 (28%)
MeSH	10,637 (40%)
SNOMED Int.	15,831 (14%)
WHO-ART	1,392 (81%)

Table 13. Number of terms from each terminology having exact correspondence (conceptual approach).

7.2 Lexical approach

There are 266,139 correspondences exist between at least two terms of the HMTP (English and French). However, the majority of correspondences have not yet been evaluated. Terminologies included in the HMTP in English and French were aligned using the two lexical approaches. Table 14 displays a fragment of the entire matrix mapping between all terminologies of the HMTP. For example, the MeSH, SNOMED International, ORPHANET and ATC terminologies were aligned using English and French lexical approaches. However, some terminologies were mapped using an English (SNOMED CT, PSIP Taxonomy) or French (CISMeF, DRC) lexical approach alone. All exact correspondences were integrated into the HMTP (Figure 2).

8. Use of alignments

8.1 Alignments for information retrieval

8.1.1 Information retrieval

Thanks to the multiple inter and intra terminology relations derived, the information retrieval results can be improved and can better respond to user's queries through "query expansion"

Fig. 1. The three types of conceptual approach integrated into the HMTP (Example of the MedDRA term "Disorientation").

or "query reformulation". Inter and Intra relations will be used to ensure navigation between terminologies. Thus, we can find all the possible connections between the terms of query in a given terminology and all other terms in other terminologies. This process can widen the scope of the search for the user according to its context without impacting the relevance of the information or the precision of the system. For example, according to the mapping between the MeSH term "Hearing aids" and the SNOMED Int term "Auditory system" we can expand the results and return all resources indexed by both terms.

8.1.2 InfoRoute

InfoRoute (Darmoni et al., 2008) is a French Infobutton (Cimino et al., 1997) developed by CISMeF. It allows the search of the main institutional websites to access high-quality documents available in French on the Internet. The CISMeF team selected fifty websites produced by high-quality Internet publishers (Figure 3), such as governments from French-speaking countries (France, Switzerland, Belgium, Canada and many African countries), national health agencies, medical societies and medical schools. Health documents on the Internet may be accessed through their description with the MeSH thesaurus: MEDLINE bibliographic database, French CISMeF, Australian Healthinsite, UK Intute catalogs. Therefore, the use of correspondences between MeSH and all terminologies used

	FMA	MedDRA	MeSH	ORPHANET	SNOMED Int	WHO-ART
CCAM	0	110	305	0	430	5
CISMeF	9	99	517	11	222	17
CISP2	7	138	219	30	254	109
Codes for drugs	0	24	1,455	3	302	0
FMA		119	1,745	32	5,777	3
ICD10	10,209	2,380	3,827	947	7,474	1,134
MedDRA	119		3,728	885	5,360	1,278
MEDLINEPlus	34	314	675	138	448	170
MeSH	1,745	3,728		1,805	15,127	1,417
ORPHANET	32	885	1,805		1,635	284
SNOMED Int	5,777	5,360	15,127	1,635		1,747
WHO-ART	3	1,278	1,417	284	1,747	
WHO-ATC	61	58	3,533	0	1,581	4
WHO-ICF	178	9	294	2	222	7
WHO-ICPS	1	13	159	0	114	6

Table 14. Fragment of the entire matrix mapping from HMTP.

to index documents in these websites is a good solution. For example: MeSH to ORPHANET (ORPHANET website), MeSH to MEDLINEplus Topics (MEDLINEPlus).

8.2 Alignments for translation

Methods developed to align biomedical terminologies were also used to translate automatically several biomedical terminologies. For example in (Deléger et al., 2010) we have combined the UMLS-based approach (conceptual approach) and a corpus-based approach to translate MEDLINEPlus® Topics from English into French. The first method based on the conceptual approach brought translations for 611 terms (from 848 MEDLINEPlus PT), 67% of which were considered valid. In (Merabti et al., 2011), we have compared two methods to translate the FMA terms into French. The first one used the conceptual approach based on conceptual information from the UMLS Metathesaurus. The two approaches allowed semi-automatic translation of 3,776 FMA terms from English into French, which was added to the existing 10,844 French FMA terms in the HMTP (4,436 FMA French terms and 6,408 FMA terms manually translated). The same approaches were used to translate 114,917 SNOMED CT English terms (40%) to at least one French term. For the FMA translation for example, evaluation methods demonstrated that 59% of the translations were rated as "good" for lexical approach and 69% for the conceptual approach. These approaches are integrated into the HMTP to translate automatically English terms to French. However, to improve the quality of the trans-lation a manual validation is needed in parallel of this automatic processing.

9. Discussion

In this chapter, we have presented the problem of integrating heterogeneous sources of medical terminologies such as thesauri, classifications, nomenclatures or controlled vocabularies to allow semantic interoperability between systems. Terminology alignment is the task of creating links between two original terminologies. These links could be

Fig. 2. Mapping of the MeSH term "myocardial infarction" according to the lexical approach in HMTP (Exact correspondence).

Fig. 3. CISMeF InfoRoute.

equivalences, correspondences or relations between terms and concepts having the same meaning but expressed with different labels. We have also presented the main methods that are commonly used for alignment between ontologies and show how we have derived them for medical terminologies. Structural methods are independent of language but the lexical ones we have presented stand for medical terminologies expressed in English and French. We have also proposed a method for evaluating sets of correspondences. All the sets of correspondences and relations we have derived are used in different contexts of information retrieval through the CISMeF catalogue and accessed through the Health Multiple Teminologies Portal developed at Rouen University Hospital. The essential difference between the alignments included in the HMTP and BioPortal (Ghazvinian et al., 2009a) is that the latter has applied lexical matching of preferred names and synonyms in English to generate alignments between concepts in BioPortal ontologies. Thus, they may miss a connection between two ontologies that actually have a significant amount of overlap in terms of the actual concepts they represent simply because these concepts have different lexical structures in the two ontologies. However, users can browse the correspondences, create new correspondences, upload correspondences created with other tools, download the correspondences stored in BioPortal, or comment on them and discuss them. Many works on aligning medical terminologies have been published recently showing that it is an active research area. In (Alecu et al., 2006), when mapping MedDRA to SNOMED CT instead of considering an unmapped MedDRA term, they considered its mapped ancestor by exploiting hierarchical relations (structure level approach). In (Bodenreider, 2009) when mapping SNOMED CT to MedDRA hierarchical relations from SNOMED CT, which are far more fine-grained than those from MedDRA, were exploited and enabled on the whole over 100 000 new mappings. However these two studies attempted to find correspondences of MedDRA terms as such, without completing the approach from a lexical standpoint trying for example to decompose and then align them to more than one SNOMED CT term. Indeed, in (Ghazvinian et al., 2009b) the comparison of different alignment approaches for medical terminologies shows that simple lexical methods perform best since medical terminologies have strongly controlled vocabularies and share little structure. Finally a specific browser was designed in order to align frequent MedDRA terms with SNOMED CT terms (Nadkarni & Darer, 2010). It was enriched with simple synonyms from the UMLS and considered decompositions of MedDRA terms. In (Diosan et al., 2009) the authors propose an automatic method for aligning different definitions taken from general dictionaries that could be associated with the same medical term although they may have the same label. The terms are those included in the CISMeF database. The method used is based on classification by Support Vector Machines derived from methods for aligning sentences from bilingual corpora (Moore, 2002). In (Milicic Brandt et al., 2011) the authors present a similar method for creating mappings between the ORPHANET thesaurus of rare diseases and the UMLS, mainly for aligning it with SNOMED CT, the MeSH thesaurus and MedDRA. The authors also use the lexical tool Norm included in the UMLS Lexical Tools to normalize terminologies included in the UMLS and normalize the ORPHANET thesaurus by "aggressive" normalization adding more steps in the process for example removing further stop words such as "disease" or "disorder". In (Mougin et al., 2011) the authors present a method for mapping MedDRA and SNOMED CT via the UMLS. They propose an automatic lexical-based approach with normalization, segmentation and tokenization steps. This approach is completed by filtering terms according to the UMLS Semantic Network: if mapping is exact but the terms do not belong to the same Semantic Type, the resulting

mapping is eliminated from the sets of mappings to be evaluated. However, this method of filtering cannot be applied when a terminology is not included in the UMLS. The evaluation in this study is quantitative and qualitative and the aim was to explore adverse drug reactions in clinical reports. Nonetheless, these correspondences are not used in concrete applications that propose semantic interoperability between systems such as the HMTP.

10. Conclusion

To summarize, we were able to achieve automatic alignment between Biomedical Terminologies. The methods we have proposed be applied to map English or French terms. The results obtained through these methods differ according to type of terminology and number of target terms used to map the source terminology. These methods are also used to translate some English terminologies to French (SNOMED CT, MEDLINEPlus).

11. Acknowledgements and funding

12. References

Alecu, I., Bousquet, C., Mougin, F. & Jaulent, M. (2006). Mapping of the whoŰart terminology on snomed ct to improve grouping of related adverse drug reactions, *Stud Health Inform* 124: 833–838.

Aronson, A. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program, *AMIA Annu Symp Proc*, pp. 17–21.

Aymé, S., Urbero, B., Oziel, D., Lacouturier, E. & Biscarat, A. (1998). Information on rare diseases: the ORPHANET project, *Rev Med Interne* 19(Suppl 3): 376S–377S.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology, *Nucleic Acids Res* 32: 267–270.

Bodenreider, O. (2009). Using snomed ct in combination with meddra for reporting signal detection and adverse drug reactions reporting, *AMIA Ann Symp Proceedings*.

Bodenreider, O. & McCray, A. T. (1998). From french vocabulary to the unified medical language system: a preliminary study., *Stud Health Technol Inform* 52 Pt 1: 670–674.

Bodenreider, O., Nelson, S. J., Hole, W. T. & Chang, H. F. (1998). Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies, *Proc. AMIA Symp. 1998*, pp. 815–819.

Brown, E., Wood, L. & Wood, S. (1999). The Medical Dictionary for Regulatory Activities (MedDRA), *Drug Safety* 2: 109–117.

Browne, A., G, D., Aronson, A. & AT, M. (2003). Umls language and vocabulary tools, *AMIA Annu Symp Proc*, p. 798.

Chute, C., Elkin, P., Sheretz, D. & Tuttle, M. (1999). Desiderata for a clinical terminology server, *Proc. AMIA Symp. 1999*, pp. 42–6.

Cimino, J., Elhanan, G. & Zeng, Q. (1997). Supporting Infobuttons with terminoligical knowledge, *JAMIA* 4(Suppl): 528–532.

Cormont, S., Vandenbussche, P., Buemi, A., Delahousse, J., Lepage, E. & Charlet, J. (2011). Implementation of a platform dedicated to biomedical analysis terminologies management, *Proc. AMIA Annu Symp Proc*.

Cornet, R. & de Keizer, N. (2008). Forty years of SNOMED: a literature review, *BMC Medical Informatics and Decision Making* 8(Suppl 1): S2.

Côté, R. A., Rothwell, D. J., Patolay, J., Beckett, R. & Brochu, L. (1993). The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International.

Darmoni, S. J., Joubert, M., Dahamna, B., Delahousse, J. & Fieschi, M. (2009a). SMTS: a French Health Multi-Terminology Server, p. 808.

Darmoni, S. J., Pereira, S., Névéol, A., Massari, P., Dahamna, B., Letord, C., Kerdelhué, G., Piot, J., Derville, A. & Thirion, B. (2008). French infobutton: an academic and business perspective., *AMIA Annu Symp Proc* p. 920.

Darmoni, S., Sakji, S., Pereira, S., Merabti, T., Prieur, E., Joubert, M. & Thirion, B. (2009b). Multiple terminologies in an health portal: automatic indexing and information retrieval, *Artificial Intelligence in Medicine*, Lecture Notes in Computer Science, Springer, Verona, Italy, pp. 255–259.

David, J. & Euzenat, J. (2008). On fixing semantic alignement evaluation measures, *The Third International Workshop on Ontology Matching*.

Deléger, L., Merabti, T., Lecroq, T., Joubert, M., Zweigenbaum, P. & Darmoni, S. (2010). A Twofold Strategy for Translation a Medical Terminology into French, *Proc. AMIA Symp. 2010*, pp. 152–6.

Diosan, L., Rogozan, A. & Pécuchet, J. (2009). *Automatic Alignment of Medical Terminologies with General Dictionaries for an Efficient Information Retrieval*, IGI Publisher, chapter 5, pp. 78–105.

Doan, A., Noy, N. & Halvey, A. (2004). Introduction to the special issue on semantic integration, *SIGMOD Record* 33: 11–13.

Douyère, M., Soualmia, L. F., Névéol, A., Rogozan, A., Dahamna, B., Leroy, J.-P., Thirion, B. & Darmoni, S. J. (2004). Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway., *Health Info Libr J* 21(4): 253–261.

Ehrig, M. & Euzenat, J. (2005). Relaxed precision and recall for ontology matching, *Integrating Ontologies Workshop*.

Euzenat, J. (2007). Semantic precision and recall for ontology alignement evaluation, *IJCAI*.

Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P. & Trojahn, C. (2011). Ontology alignment evaluation initiative: six years of experience, *Journal on Data Semantics* .

Euzenat, J. & Shvaiko, P. (2007). *Ontology Matching*, Hiedelberg: Springer-Varlag.

Fellbaum, C. (ed.) (1998). *WordNet: an electronic lexical database*, MIT Press.

Ferru, P. & Kandel, O. (2003). Dictionnaire des résultats de consultation (révision 2003-04), *Doc Rech Med Gen* 62: 3–54.

Fung, K. & Bodenreider, O. (2005). Utilizing UMLS for semantic mapping between terminologies, *Proc AMIA Symp*, pp. 266–270.

Ghazvinian, A., Noy, N., Jonquet, C., Shah, N. & Musen, M. (2009a). What four million mappings can tell you about two hundred ontologies, *Proceedings of the 8th ISWC*.

Ghazvinian, A., Noy, N. & Musen, M. (2009b). Creating mappings for ontologies in biomedicine: Simple methods work, *AMIA Annual Symposium*.

Granitzer, M., Sabol, V., Onn, K., Lukose, D. & Tochtermann, K. (2010). Ontology alignement-A Survey with focus on visually supported semi-automatic techniques, *Future Internet* 2(3): 238–258.

Grosjean, J., Merabti, T., Dahamna, B., Kergouraly, I. & Thirion, B. (2011). Health Multi-Terminology Portal: a semantic added-value for partient safety, *Patient Safety Informatics - Adverse Drug Events, Human Factors and IT*, Vol. 166, pp. 129–138.

Gruber, T. (1993). Toward principles for the design of ontologies used for knowledge sharing, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Kluwar Academic Publishers.

Hamming, R. (1950). Error detecting and error correcting codes, *Technical report*, Bell System Technical Journal.

Imel, M. (2002). A closer look: the SNOMED clinical terms to ICD-9-CM mapping., *J AHIMA* 73(6): 66–9; quiz 71–2.

Jaccard, J. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines, *Bulletin de la société vaudoise des sciences naturelles* 37: 241–272.

Johnson, H., Cohen, K., Baumgartner, W., Lu, Z., Bada, M., Kester, T., Kim, H. & Hunter, L. (2006). Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies, *Pacific Symposium on Biocomputing*.

Joubert, M., Abdoune, H., Merabti, T., Darmoni, S. & Fieschi, M. (2009). Assisting the translation of SNOMED CT into French using UMLS and four representative French-language terminologies, *Proc. AMIA Symp.*, pp. 291–295.

Kotis, K. & Lanzenberger, M. (2008). Ontology matching: current status, dilemmas and future challenges, *International Conference on Complex, Intelligent and Software Intensive Systems*, pp. 924–927.

Lamy, J.-B., Ducols, C., Bar-Hen, A., Ouvrard, P. & Venot, A. (2008). An iconic language for the graphical representation of medical concepts, *BMC Bioinformatics* 8: 16.

Lefevre, P. (2000). *La recherche d'informations : du texte intégral au thésaurus*, Editions Hermès.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals, *Sov. Phys. Dokl.* pp. 707–710.

Lin, D. (1998). An information-theoretic definition of similarity, *Proc. Int. Conf. on Machine Learning*, pp. 296–304.

Lindberg, D. A., Humphreys, B. L. & McCray, A. T. (1993). The unified medical language system., *Methods Inf Med* 32(4): 281–291.

McCray, A., Srinivasan, S. & Brown, A. (1994). Lexical methods for managing variation in biomedical terminologies, *Annual Symposium on Computer Applications in Medical Care*, pp. 235–239.

Merabti, T. (2010). *Methods to map health terminologies: contribution to the semantic interoperability between health terminologies*, PhD thesis, University of Rouen.

Merabti, T., Joubert, M., Lecroq, T., Rath, A. & Darmoni, S. (2010a). Mapping biomedical terminologies using natural language processing tools and UMLS: mapping the Orphanet thesaurus to the MeSH, *Biomedical Engineering and Research* 31(4): 221–225.

Merabti, T., Massari, P., Joubert, M., Sadou, E., Lecroq, T., Abdoune, H., Rodrigues, J. & Darmoni, S. (2010b). Automated approach to map a French terminology to UMLS, *MedInfo2010*, Vol. 160, Cap Town, South Africa, pp. 1040–1044.

Merabti, T., Pereira, S., Letord, C., Lecroq, T., Dahamna, B., Joubert, M. & Darmoni, S. (2008). Searching related resources in a quality controlled health gateway: a feasibility study, *The XXIst International Congress of the European Federation for Medical Informatics (MIE'08)*, Vol. 136, pp. 235–249.

Merabti, T., Soualmia, L. F., Grosjean, J., Palombi, O., Müller, J.-M. & Darmoni, S. J. (2011). Translating the Foundational Model of Anatomy into French using knowledge-based and lexical methods, *BMC Medical Informatics and Decision Making* 11: 65.

Milicic Brandt, M., Rath, A., A., D. & S., A. (2011). Mapping orphanet terminology to umls, *Proceedings of 13th Conference on Artificial Intelligence in MEdicine, AIME*, p. 194Ű203.

Miller, N., Lacroix, E. M. & Backus, J. E. (2000). MEDLINEplus: building and maintaining the national library of medecine's consumer health web service, *Bull Med Libr Assoc* 88(1): 11–7.

Moore, R. (2002). Machine translation: From research to real users, *Proceedings of ATMA, Lecture Notes in Computer Science*, Vol. 2499, pp. 135–144.

Mougin, F., Dupuch, M. & Grabar, N. (2011). Improving the mapping between meddra and snomed ct, *Proceedings of 13th Conference on Artificial Intelligence in MEdicine*.

Nadkarni, P. & Darer, J. (2010). Determining correspondences between high frequency meddra concepts and snomed: a case study, *BMC Med Infor. Decis. Mak* 10: 66.

Nelson, S., Johnston, D. & Humphreys, B. (2001). *Relationships in Medical Subject Headings*, New York: Kluwer Academic Publishers, pp. 171–184.

Noy, N., Musen, M., Mejino, J. L. & Rosse, C. (2004). Pushing the envelope: challenges in a frame-based representation, *Data Knowl. Eng.* 48: 335–359.

Pereira, S. (2007). *Muti-Terminology indexing of concepts in health*, PhD thesis, University of Rouen.

Pereira, S., Névéol, A., Kerdelhué, G., Serrot, E., Joubert, M. & Darmoni, S. J. (2008). Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue., *AMIA Annu Symp Proc* pp. 586–590.

Peters, L., Kapusnik-Uner, J. & Bodenreider, O. (2010). Methods for managing variation in clinical drug names, *Proc Annu Symp AMIA 2010*, pp. 637–4.

Rector, A., Bechhover, S. & Goble, C. (1997). The GRAIL concept modelling language for medical terminology, *Artif Intell Med* 9(2): 139–71.

Robinson, P. & Mundlos, S. (2010). The human phenotype ontology, *Clin Genet* 77: 525–534.

Roche, C. (2005). Terminologie et ontologie, *Revue Langages* 157: 48–62.

Rodrigues, J., Trombert-Paviot, B., Baud, R., Wagner, J. & Meusnier-Carriot, F. (1997). Galen-in-use : Using artificial intelligence terminology tools to improve the linguistic coherence of a national coding system for surgical procedures, *Proceedings of the 15th International Congress of the European Federation for Medical Informatics, MIE-97*, p. 897Ű901.

Rosse, C. & Mejino, J. (2003). A reference ontology for biomedical informatics: the Foundational Model of Anatomy, *Journal of Biomedical Informatics* 36: 478–500.

Sakji, S., Lethord, C., Pereira, S., Dahamna, B., Joubert, M. & Darmoni, S. J. (2009). Drug information portal in Europe: Informatio retrieval with multiple health terminologies, *Stud Health Technol Inform*, Vol. 150, pp. 497–501.

Salton, G. & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

Sarker, I., Cantor, M., Gelman, R., Hartel, F. & Lussier, Y. (2003). Linking biomedical language information and knowledge ressources in the 21st Century: GO and UMLS, *Pacific Symposium on BioComputing*, Vol. 8, pp. 439–450.

Shvaiko, P. & Euzenat, J. (2005). A survey of schema-based matching approaches, *Journal on Data Semantics IV* pp. 146–171.

Soualmia, L. (2004). *Towards an intelligent IR through the evaluation of multiple knowledge-based methods of query expansion: application to the health domain on the Internet*, PhD thesis, INSA Roue.

Soualmia, L. F., Griffon, N., Grosjean, J. & Darmoni, S. J. (2011). Improving Information Retrieval by Meta-modelling Medical Terminologies, *Proceedings of 13th Conference on Artificial Intelligence in MEdicine, AIME*, Vol. 6747 of *Lectures Notes in Computer Science*, Springer, pp. 215–219.

Stoilos, G., Stamou, G. & Kollias, S. (2005). A string metric for ontology alignement, *International Semantic Web Conference*, Vol. 3729, pp. 624–637.

Wang, Y., Patrick, J., Miller, G. & O'Hallaran, J. (2008). A computational linguistics motivated mapping of ICPC-2 PLUS to SNOMED CT, *BMC Med Inform Decis Mak* 8 Suppl 1: 5.

Wu, Z. & Palmer, M. (1994). Verb semantics and lexical selection, *32nd Annual Meetings of the Associations for Computational Linguistics*, pp. 133–138.

Zeng, M. & Chan, L. (2004). Trends and issues in establishing interoperability among knowledge organization systems, *Journal of the American Society for Information* 55: 377–395.