Manuscript Number: IRBM-D-09-00044R1

Title: Mapping biomedical terminologies using natural language processing tools and UMLS: mapping the Orphanet thesaurus to the MeSH

Article Type: Article original

Keywords: CIM10; France; traitement langage naturel; vocabulaire contrôlé; UMLS;

Corresponding Author: Prof. Stefan J. Darmoni, MD, PhD

Corresponding Author's Institution: Rouen University HOspital

First Author: Stefan J. Darmoni, MD, PhD

Order of Authors: Stefan J. Darmoni, MD, PhD; Merabti Tayeb, MS; Joubert Michel, PhD; Tayeb Merabti, MS; Michel Joubert, PhD; Thierry Lecroq, PhD; Ana Rath, MD

Abstract: Background: Orphanet aims to provide rare disease information to healthcare professionals, patients, and their relatives. Objective: The objective of this work is to evaluate two methodologies (UMLS and manual Orphanet-ICD-10 link-based mapping & String Based matching) used to map Orphanet thesaurus to the MeSH thesaurus. Results: On a corpus of 375 mappings, the string based matching provides significantly better results than the UMLS and manual Orphanet-ICD-10 link-based mapping. Conclusion: String based matching could be applied to any biomedical terminology in French not yet included into UMLS.

Mapping biomedical terminologies using natural language processing tools and UMLS: mapping the Orphanet thesaurus to the MeSH

Tayeb Merabti, MS[1, 2], Michel Joubert, PhD[2], Thierry Lecroq[1], PhD,
A. Rath, MD[3], Stefan J. Darmoni, MD, PhD[1]
[1]CISMeF, University Hospital, Rouen, France & TIBS, **LITIS EA 4108, Institute of Biomedical Research, University of Rouen, France***; [2]LERTIM EA 3283, Faculty of Medicine, Marseille, France; [3] Orphanet INSERM SC11, Paris, France*

Abstract

Background: *Orphanet aims to provide rare disease information to healthcare professionals, patients, and their relatives.* **Objective**: *The objective of this work is to evaluate two methodologies (UMLS and manual Orphanet-ICD-10 link-based mapping & String Based matching) used to map Orphanet thesaurus to the MeSH thesaurus.* **Results**: On a corpus of 375 mappings, *the string based matching provides significantly better results than the UMLS and manual Orphanet-ICD-10 link-based mapping.* **Conclusion**: *String based matching could be applied to any biomedical terminology in French not yet included into UMLS.*

Introduction

Retrieval and exchange of information and data from multiple health terminologies and databases is desirable. Interoperability is intended to facilitate the coding, transmission and use of medical concepts across a wide spectrum of health actors. However, translating information from one terminology to another is not very easy because of the heterogeneity of health terminologies.

The process of terminology mapping consists of identifying identical concepts or relationships between terminologies [1]. A number of algorithms and approaches have been proposed to create an automatic mapping between health terminologies [1-5]. For example, Rocha et al [2] and Cimino et al. [3] both proposed a frame-based approach to perform mappings between health terminologies. Other approaches were proposed using UMLS (Unified Medical Languages Systems) [6] as a knowledge resource to perform mappings between terminologies. For example, Fung and Bodenreider [4] described an algorithm [5] to map between any two terminologies in the UMLS making use of synonymy, explicit mapping relations and hierarchical relationships. However, approaches using UMLS are limited to the biomedical terminologies already incorporated into UMLS.

The objective of this work is to present a mapping method to be used by any biomedical terminology in French not yet included in the UMLS to be included in this metathesaurus. We proposed two different mapping strategies to map Orphanet nomenclature of rare diseases [7], a biomedical terminology on rare diseases not yet included into the UMLS to the MeSH (Medical Subject Heading) thesaurus included into the UMLS.

Material

The first strategy of this work uses UMLS and an external manual mapping of Orphanet terms to ICD-10 code. The second strategy uses only natural language processing without using UMLS to make a direct and an automatic mapping between Orphanet and MeSH. We also provide an evaluation and a comparison of these two strategies for this mapping.

The MeSH thesaurus was chosen as the target terminology for comparing mapping strategies for two main reasons:

1. 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 term (e.g. Marfan syndrome) and one PubMed query. The CISMeF team has a strong experience with the MeSH thesaurus. Therefore, the evaluation will be using one CISMeF expert.

2. The MeSH is the second largest terminology available represented into UMLS with 24,767 preferred terms (after SNOMED) versus 17,867 for MedDRA in French for example.

Nevertheless, the method presented is easily extendable to all the health terminologies from the UMLS, which has a French translation.

This work is a part of the European Union, DG SANCO, “Scientific Support to the Rare Disease Task Force Activities” project and it is done in collaboration with Orphanet. The overall project is the mapping of the multi-lingual Orphanet thesaurus into all French terminologies included into the “Health Multi-Terminological Server” [8]. Therefore, the Orphanet mapping to UMLS is an imperative task to be performed.

ICD-10: the “International Classification of Diseases” [9] is designed to promote international comparability in the collection, processing, classification, and presentation of mortality and morbidity statistics. These include the analysis of the general health situation of population groups and monitoring of the incidence and prevalence of diseases and other health problems in relation to other variables such as the characteristics and circumstances of the individuals affected, reimbursement, resource allocation, quality and guidelines. It is published by the World Health Organization. ICD is revised periodically and is currently at its tenth edition launched in 1992. An ICD-11 version is planned for 2014.

Orphanet: *Orphanet* aims to provide rare disease information to healthcare professionals, patients, and their relatives, in order to contribute to the improving of the diagnosis, care and treatment of these diseases. Orphanet has also developed a multi-hierarchical thesaurus for rare diseases, available in five European languages (English, French, Spanish, German & Italian) (including 7,427 entries and 4259 synonyms). WHO planned to integrate this Orphanet thesaurus in the ICD version 11 currently in progress.

UMLS: the “Unified Medical Language System” is a repository of biomedical vocabularies developed by the US National Library of Medicine. Currently, the UMLS integrates over 5 million names for over 1,270,000 concepts from more than 140 biomedical terminologies, classifications, and ontologies, as well as 13 million relations among these concepts.

The UMLS is made up of three main knowledge components, but, for our purpose, we retain only the Metathesaurus: a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. Each concept isolated from terminologies has a concept unique identifier (CUI) in the Metathesaurus. This means that the same concept appearing in various terminologies, perhaps with various names and synonyms, has a unique entry in the Metathesaurus.

MeSH: the National Library of Medicine's controlled vocabulary thesaurus originally intended to index scientific articles for the Index Medicus and for the MEDLINE database. There are 24,767 descriptors and 83 qualifiers (and subheadings) in the 2008 version. There are also over 97,000 entry terms (or synonyms) that assist in finding the most appropriate MeSH Headings. Since 1993, the CISMeF team is adaptating the MeSH thesaurus to index health French resources instead of scientific articles [10]. The CISMeF team has added over 10, 000 French synonyms (including acronyms).

Methods

Strategy 1: UMLS and manual Orphanet-ICD-10 link-based mapping

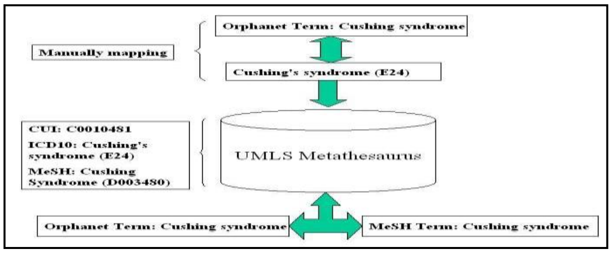
This strategy is based on the external manual mapping between Orphanet and ICD-10 terms performed by Orphanet. There are 2,083 Orphanet terms manually mapped to at least one ICD-10 code (28% of all Orphanet terms).

In this approach, the link provided by the UMLS metathesaurus between ICD-10 and MeSH was used. The 2008AA version of the UMLS was used. It contains the MeSH 2008 version and ICD-10. Hence, an effective mapping exits between two in ICD-10 and MeSH if these terms share the same UMLS Concept Unique Identifier (CUI) in the metathesaurus. For example, there is an effective mapping between the ICD-10 term “Cushing syndrome” (Code: E24) and the MeSH term “Cushing syndrome” since they shared the same UMLS Concept CUI: C0010481).

Table 1. Example of UMLS and manual Orphanet-ICD-10 links based mapping

Orphanet Term	ICD-10 term	MeSH term
Cushing syndrome	Cushing's syndrome	Cushing Syndrome
Ichthyosis, X-linked	X-linked ichthyosis	Ichthyoses, X-Linked
Muscular dystrophy, Duchenne and Becker types	Muscular dystrophy	Muscular Dystophies

Figure 1. Mapping Schema with UMLS


Orphanet Term: Cushing syndrome
Manually mapping
Cushing's syndrome (E24)
CUI: C0010481
ICD10: Cushing's syndrome (E24)
MeSH: Cushing Syndrome (D003480)
UMLS Metathesaurus
Orphanet Term: Cushing syndrome
MeSH Term: Cushing syndrome

Strategy 2: String Based matching

In this strategy, a “string based matching” method was used to link directly terms of Orphanet and MeSH without using UMLS. This method allows from a given term in the source terminology (Orphanet) to find a term in the target terminology (MeSH) that is the most lexically similar. Two terms are considered as lexically similar if all the words composing the two terms are the same. For example, the Orphanet term “LCAT deficiency” is lexically similar to the MeSH term “lcat deficiency” because the two words “lcat” and “deficiency” compose both Orphanet and MeSH terms. According to its definition, the “String based matching” cannot take into account inflections, stop-words, etc.

To overcome this problem, basic natural language processing is necessary beforehand: (a) remove stop words: frequent short words that do not affect the phrases such as “a”, “Nos”, “of” etc. are removed from all terms in both terminologies; (b) stemming: we use a French stemming “Lucene” which proved to be the most efficient for the F-MTI automatic indexing tools using several health terminologies [11], as compared to the stemming tool developed by the CISMeF team & the stemming tool [12].

This approach was applied for all the MeSH terms including preferred and synonyms terms, since there are several terms in Orphanet that correspond to MeSH synonyms terms if string based matching is used. For example, the Orphanet term “Disease of Kimura” corresponds to the MeSH term “Angiolymphoid Hyperplasia Eosinophilia” since it has the synonym term “Disease, Kimura”.

Table 2. Example of string based matching

Orphanet Term	MeSH Term
Hypokalemic periodic paralysis	Hypokalemic Periodic, Paralysis
West syndrome	Spasms, Infantile
Kimura disease	Angiolymphoid with Hyperplasia Eosinophilia

Evaluation & comparison

To evaluate the two methods four sets of mapping results were created from the results of the two methods applied to only 2,083 Orphanet terms manually mapped to ICD-10:

1. **First set:** The mapping results achieved by the first strategy “UMLS and manual Orphanet-ICD-10 link-based mapping” and not by the second “String Based matching” (only manually found)

2. **Second set:** The mapping results found by the second strategy and not by the first (Only string based matching found)

3. **Third set:** The discrepant mapping results found by both strategies for the same Orphanet term. For example, for the Orphanet term “Tangier disease” the two strategies found two different MeSH terms, the MeSH term “Hypolipoprotenemia” with the first strategy and the MeSH term “Tangier disease” with the second.

4. **Fourth set:** The mapping results found with both strategies (the same mapping results)

Four evaluations were done by a physician (SJD), head of the CISMeF team [13] according to the four sets of mapping results obtained. The evaluation was made on only 100 mapping results of each set, which were randomly obtained. The following terms were used to describe the quality of each mapping result: “relevant” the mapping between one MeSH term and one Orphanet term was rated as correct; “non-relevant” when the mapping between MeSH and Orphanet terms was considered by the expert as not correct; “BT-NT” the Orphanet term was rated as broader than the MeSH corresponding term; “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”; “Sibling” when the MeSH corresponding and Orphanet term are siblings (from the MeSH point of view). For example, “Cryptophthalmia, isolated” is evaluated as the sibiling of “microphthalmos”. A statistical comparison was performed for each set of mapping results. To this end, a Fisher 's exact test was used to compare the two methods.

Results

UMLS and manual Orphanet-ICD-10 link-based mapping:

Among the 2,083 Orphanet terms (28% of all Orphanet terms) manually mapped to at least one ICD-10 code, 619 possible matchings were found to at least one MeSH terms using UMLS (30% from 2,083).

String Based matching:

Among the 2,083 Orphanet terms linked manually to at least one ICD-10 code, 593 possible matching were found to at least one MeSH terms (28% from 2,083). However, 1,004 possible matchings were done to at least one MeSH term (13% from 7,424) when this method was applied to all Orphanet terms. According to the results of each method we obtained:

1. **First set:** 327 mapping results were found only by the “UMLS and manual Orphanet-ICD-10 manual mappings” and not by the “string based matching mapping”

2. **Second set:** 306 mapping results were found only by the “string based matching mapping”

3. **Third set:** 75 different mapping results were found by both methods with the same Orphanet terms.

4. **Fourth set:** 211 same mapping results were founded by both methods.

The evaluation of mapping results obtained by each strategy independently is displayed in table 3. Overall 85% of mapping results obtained by the strategy 2 (String Based matching) are ranked as relevant when only 21 % of mapping results are ranked as relevant for the first strategy (UMLS and manual Orphanet-ICD-10 link-based mapping), whereas 32% and 15 % of the mapping results obtained by strategy 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 3. Evaluation results of the two sets of mapping results (Mapping results found by each strategy only)

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

*The Fisher's exact test for the Table 3 gives significant results (p <0.00001).

Table 4 displays the evaluation results for the third set corresponding to different mapping results by the two strategies for the same Orphanet term. For the first strategy (UMLS and manual Orphanet-ICD-10 link-based mapping), overall 39 mapping results are evaluated as "BT-NT" when only 6 mapping results are evaluated as "relevant". For the second strategy (String Based matching), there are overall 62 mappings results evaluated as "relevant", whereas there are 8 mapping results evaluated as "BT-NT".

Table 4. Evaluation results of the fourth set of mapping results (For the same Orphanet term different results mapping)

	Relevant	BT-NT	NT-BT	Sibling	Non-relevant
UMLS and manual Orphanet-ICD-10 link-based mapping	6	39	7	2	21
String Based matching	62	8	1	1	2

The Fisher's exact test for the Table 4 gives significant results (p <0.00001).

The results of evaluation for the fourth set corresponding to the same mapping results founded by each strategy found relevant mapping in 98% cases and BT-NT relations in 2% cases.

Table 5 displays some examples for each 5 types of evaluations performed.

Table 5. Example of each type of evaluation performed

Type of evaluation	Orphanet term	Target term (MeSH)
Relevant	Nocardiosis	Nocardia infections
BT-NT	Hemophilia	Hemophilia a
NT-BT	Duchenne and Becker muscular dystrophy	muscular dystrophies
Sibiling	Cryptophthalmia, isolated	microphthalmos
Non-relevant	Disorder of sex development	Pseudohermaphroditism

Discussion

Quantitatively, on the same corpus of 2,083 Orphanet terms which are manually mapped to ICD-10, the first strategy (UMLS and manual Orphanet-ICD-10 link-based mapping) is providing a little more mapping than the second strategy (String Based matching): 619 vs. 593. But the second strategy is also providing 511 additional mappings for Orphanet terms, which cannot be manually mapped, to the ICD-10 (by the first strategy).

Qualitatively, the second strategy (String Based matching) is providing significantly better results than the first strategy (UMLS and manual Orphanet-ICD-10 link-based mapping) (Fisher's exact test -p <0.00001- on Tables 3 & 4). Furthermore, the String Based matching strategy does not require any explicit manual mapping, which is a highly time consuming task.

This difference could be explained in part by the underrepresentativeness of rare diseases in the current ICD-10. In fact, only 240 rares diseases have a specific code in ICD-10 whereas 259 of them are just included in an ICD-10 category (AR personal communication). Comparison of different coding systems using ICD-10 for coding rare diseases shows that discrepancies amongst coders are due to ambiguity of the ICD-10 or to the need for human interpretation when coding (AR personal communication). The first mapping strategy would achieve different results when applied to other datasets for ICD-10 coding of rare diseases is coder-dependent in ICD-10 current version. This is a further reason to have automated mapping strategies based on lexical and semantic criteria.

The “string based matching mapping” could be applied to any health terminology available in French not yet included into UMLS to be included into UMLS. This method could also be extended to partial and combination string based matching, which will provide a 0/N mappings between an Orphanet terms and a MeSH term. For example, using the “combination string based matching” the Orphanet term “*Rare prediatric biliary tract disease*” will be mapped to three MeSH terms “*biliary tract*”, “*children*” and “*rare disease*”.

This work is performed as a part of the EC-funded “Scientific Support to the Rare Disease Task Force Activities”.

Acknowledgments

This work was partially granted by the European Union, DG SANCO A/101112 “Scientific Support to the Rare Disease Task Force Activities” & InterSTIS project funded by the French National Research Agency[1]. We thank also the Orphanet team to have provided the Orphanet thesaurus.

References

1. Yefeng W, Jon P, Miller G, O'Hallaran J. O'Halloran. BMC Medical Informatics & Decision Making 2008; 8(Suppl 1)**:** S5
2. Rocha RA, Rocha BH, Huff SM. Automated translation between medical vocabularies using a frame-based interlingua. Proc Annu Symp Comput Appl Med Care 1993;690-4
3. Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. MD Comput 1990;7:104-9
4. Fung KW, Bodenreider O. Utilizing UMLS for semantic mapping between terminologies. AMIA Annu Symp Proc 2005:266-70
5. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond Synonymy: exploiting the UMLS semantics in mapping vocabularies. Proceedings'/ AMIA Annual Symposium 1998; 815-9
6. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993; 32(4):281-91
7. Aymé S, Urbero B, Oziel D, Lecouturier E, Biscarat AC. Information on rare diseases: the Orphanet project Rev Med Interne. 1998; 19 Suppl 3:376S-377S.
8. Joubert M, Dahamna B, Delahousse B, Fieschi M, Darmoni SJ. SMTSR : Un Serveur Multi-Terminologies de Santé. In: Informatique & Santé, Journées Francophones d'Informatique Médicales, (in press)
9. ICD-10 (International Classification of Diseases) [http://www.who.int/classifications/icd/en/]. Accessed Feb 2009
10. Douyère M, Soualmia L, Névéol A, Rogozan A, Dahamna B, Leroy JP, et al. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Info Libr J 2004; 21(4): 253-61.
11. Pereira S. Multi-terminology indexing of concepts in health. [Indexation multiterminologique de concepts en santé]. PhD Thesis, University of Rouen, Normandy, France
12. Soualmia LF. Etude et Evaluation d’Approches Multiples d’Expansion de Requêtes pour une Recherche d’Information Intelligente : Application au domaine de la Santé sur l’Internet. PhD Thesis; December 2004.
13. Darmoni SJ, Leroy JP, Baudic F, Douyère M, Piot J, Thirion B. CISMeF: a structured health resource guide. Methods Inf Med 2000; 39(1): 30-5

Address for correspondence

Prof. Stefan Darmoni
CISMeF team, Rouen University Hospital (Cour Leschevein, Porte 21, 3ème étage)
1, rue de Germont – 76031 Rouen, FRANCE
E-mail: tayeb.merabti@chu-rouen.fr

[1] ANR-07-TECSAN-010

Réponses aux lecteurs :

Commentaire :

You just have to pay attention to the spelling (i e 'ibstead' for instead in the second page : ... French resourses instead...)

Correction :

Ibstead → instead