MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1040

An Automated Approach to map a French terminology to UMLS

Tayeb Merabti[a,b], Philipe Massari[a], Michel Joubert[b], Eric Sadou[c], Thierry Lecroq[a], Hocine Abdoune[b], Jean-Marie Rodrigues[c,d], Stefan J. Darmoni[a]

[a]*CISMeF, University Hospital, Rouen, France & TIBS, LITIS EA 4108, Institute of Biomedical Research, France*
[b]*LERTIM EA 3283, Faculty of Medicine, Marseille, France*
[c]*Department of Public Health, CHU University of Saint Etienne, France*
[d]*WHO FIC Collaborative Centre for International Classifications in French Language, Paris, France*

Abstract

Background: CCAM is a French terminology for coding clinical procedures. CCAM is a multi-hierarchical structured classification for procedures used in France for reimbursement in health care, which is external to UMLS . Objective: The objective of this work is to describe a French lexical approach allowing mapping CCAM procedures to the UMLS Metathesaurus to achieve interoperability to multiple international terminologies. This approach used a preliminary step intended to take only the significant characters used to code CCAM corresponding to anatomical and actions axes. Results: According to the 7,926 CCAM codes used in this study, 5,212 possible matches (exact matching, single to multiple matching, partial matching) are found using the French CCAM to UMLS based mapping, 65% of the corresponding anatomical terms in the CCAM code are mapped to at least one UMLS Concept and 37% of the corresponding action terms in the CCAM code are mapped to at least one UMLS Concept. For all the exact matches found (n=200), 91% were rated by a human expert as narrower than the mapped UMLS Concepts, while only 3% were irrelevant.

Keywords:

Coding system, Mapping, Ontology, Semantic interoperability, Terminology

Introduction

Retrieval and exchange of information from multiple health terminologies and databases becomes increasingly useful. UMLS appears as a powerful candidate for supporting interoperability among all biomedical terminologies. Mapping terminology for coding clinical procedures to UMLS is essential for international case mix comparison between data and practices in different Electronic Health Record systems**.** This mapping will achieve semantic interoperability between CCAM and every international terminology through UMLS, especially SNOMED International ("procedure" axis) and ICD10. All these French health terminologies are integrated into a Health Multi-Terminology server [1]. However, translating information from one terminology to another is not very easy because of their heterogeneity, due to the different scope, points of view, and level of abstraction and detail of each health terminology.

The process of terminology mapping consists of identifying identical (or approximately identical) concepts or relationships between terminologies [2]. A number of algorithms and approaches have been proposed to create an automatic mapping between health terminologies [2-6]. For example, Rocha et al [3] and Cimino et al. [4] both proposed a frame-based approach to perform mappings between health terminologies. Other approaches were proposed using UMLS (Unified Medical Languages Systems) [7] as a knowledge resource to perform mappings between terminologies. For example, Fung and Bodenreider [5] described an algorithm [6] to map between any two terminologies in the UMLS making use of synonymy, explicit mapping relations and hierarchical relationships. However, approaches using the UMLS are limited to the biomedical terminologies already incorporated into UMLS.

The objective of this work is to describe a mapping method to be used by any biomedical terminology in French not yet included in the UMLS, to be subsequently included in this metathesaurus. The mapping approach has been used and evaluated in this work to map the CCAM terminology (Classification Commune des Actes Médicaux) for procedures to UMLS Metathesaurus. This terminology is not yet included in the UMLS.

This work takes place in a more global InterSTIS project, funded by the French National Agency. Semantic interoperability inter and intra terminology is the main objective of InterSTIS. This current work was funded mainly by the InterSTIS project grant.

Materials

CCAM is a multi-hierarchical structured classification mainly for surgical procedures used in France, for reimbursement and policy making in health care. Several terminologies for procedures exist and are used in different countries. For example,

the CPT (Current Procedural Terminology) [8] developed by the American Medical Association and since 2001, selected by the Department of Health and Human Services (HHS) as the standard code set for reporting health care services in electronic transactions. The NOMESCO Classification of Surgical Procedures (NCSP) used by all the 5 Nordic countries [9].

The 10th version of the CCAM covers about 7,926 procedure codes. 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.

The concatenation of the codes for these axes results in a multiaxial code with 7 alphanumeric characters, which gives a "synthetic" procedure description, based on the code/definition tables of the CCAM Basic Coding System. (See Figure 1).

The construction of the CCAM associated the traditional expertise and ontology-driven terminological tools provided by GALEN (Generalised Architecture for Languages, Encyclopedias and Nomenclatures in Medicine) [10,11] with the constraint of being compatible with the European standards [12]. This process allowed checking the conformity of the label with the significant concepts of medical knowledge. Therefore, CCAM has been the main source of inspireation for the classification procedure in Australia (ICHI) [13] and Germany [14]

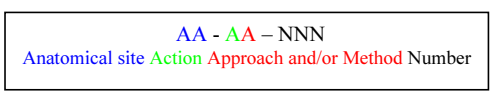

Figure 1- CCAM code structure

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 of over 1,270,000 concepts from more than 140 biomedical terminologies, classifications, and ontologies, as well as 13 million relations among these concepts. 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.

Methods

One automated mapping method was used to map CCAM codes to UMLS Concepts. This method is based on the structure of the CCAM code described in the previous section. However, it is impossible to assign one or more specific UMLS concepts using only CCAM label. This is mainly due to the length of CCAM labels. Indeed, there are 85% of CCAM labels equal or with more than 5 words vs. only 5% of the MeSH descriptors equal or with more than 5 words. The approach using the coding structure for mapping to UMLS was also used in [15] to allow the LOINC (Laboratory Observation Identifier Names and Codes) [16] integration to UMLS.

In this approach, only the significant first three characters composing the CCAM code according to the anatomic (two characters) and action (one character) axes are mapped to the UMLS Metathesaurus. For example: The CCAM code "NCCA010" with the label: "*Osteosynthesis of tibial diaphysis fracture by external fixing*", is represented according to the significant three first characters with:

- "*Bones of the leg*" corresponds to "NC" characters (anatomic axis)
- "*Osteosynthesis*" corresponds to the "C" character (action axis)

French natural languages processing tools and mapping algorithms were developed by the CISMeF team to map between French health terminologies. These tools were used in previous works [17, 18] and extended to link terms in multiple French health terminologies.

This approach allows for a given term (obtained by the concatenation of the resulting terms of the preprocess step according to the two axes (anatomic and action) to find a UMLS Concept with French terms that are lexically the most similar to it. Thus, to overcome some problems like account inflections, stop-words, etc., basic natural language processing is necessary beforehand:

1. Remove stop words: frequent short words that do not affect the phrases such as "a", "Nos", "of", etc; are removed form all terms in all terminologies (CCAM and French terminologies of the UMLS).
2. Stemming: we use a French stemmer "Lucene" which proved to be the most efficient for the F-MTI automatic indexing tools using several health terminologies [17], as compared to the stemming tools developed by the CISMeF team and the stemming tools in [18].

The mapping used by this approach provides three types of matches between all terms in source terminologies and the French terms in the UMLS metathesaurus. These levels of matching are inspired in most cases by the "ISO 5964", which is an ISO standard for the establishment and development of multilingual thesauri [19]. Relation types may be also represented in SKOS language [20]. SKOS language is also used to represent French health terminologies into the French Health Multi-terminological Server [1], which intends to integrate the main health terminologies available in French, including those not yet mapped to the UMLS (e.g. CCAM, ATC, Orphanet).

The three types of matching are:

Exact matching

One CCAM term and one French term in UMLS are in "exact matching" if all the words composing the two terms are exactly the same. Thus, according to this matching there is at most one UMLS Concept corresponding to all the significant characters of the CCAM code (see Table 1). Formally, in this type of matching the label obtained by the concatenation of the two terms according to the two axes, is considered as a one and unique term, an

"exact matching" between this term and one French term in UMLS.

Single to multiple matching

One CCAM term and at least two French terms from UMLS are in a "Single to multiple matching" when the CCAM term cannot be matched by one exactly French term in UMLS, but can be expressed by a combination of two or more French terms in UMLS. In this "Single to multiple matching", the CCAM term is mapped to at least two UMLS Concepts.

Partial matching

This type of matching is the less accurate one. In this type of matching only a part of the CCAM term will be mapped to one or more UMLS Concepts. Table 1, list some examples corresponding to the three types of matching described above.

Evaluation

The evaluation was performed on all the types of matching from the "exact" set matching type and for only 100 from the "Single to multiple" set matching type. We chose only 100 matching instances because in most cases the same codes with the same three firsts characters were mapped to the same UMLS concepts (HLHH003, HLHH004...).

The qualitative evaluation was performed by a physician (PM), expert in CCAM and in UMLS. The following terms were used to rate the quality of each matching result: (a) "equivalent" the UMLS concept corresponded exactly to the CCAM code; (b) "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; (c) "NT-BT" the CCAM code was rated as narrower than the PTs in the UMLS concept, (d) "incomplete" when the UMLS concept only reflected some part of the CCAM label and (f) "irrelevant" when the matching was considered by the expert as incorrect.

For example, the matching between the CCAM code "BFGA003" (label: "*manually extraction of lens without intraocular lens implant*") and the UMLS concept C0007389 (preferred term: "*cataract extraction*") was rated as NT-BT because the UMLS concept is narrower and less accurate 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) was evaluated independently. 2) the matching between the CCAM code and the combination of the UMLS concepts was then evaluated in this second phase. For example, to evaluate the matching between the CCAM code "AAFA003" and the two UMLS concepts: C006104 (preferred term: *"Brain") and C0919588 (*(preferred term: *"Exeresis").* First, the expert evaluated each axis with the corresponding UMLS ((Brain, C006104) =equivalent and (Exeresis, C091958) =equivalent)). Second, the expert evaluated the matching between the label and the combination of the two UMLS concepts: (AAFA003, (C006104, C091958) =NT-BT).

Table 1- Examples of the three types of matchings using the French based UMLS matching

CCAM code	Anatomic axis	Action axis	UMLS concepts	Typeof matching
BDHA001	Cornea	biopsy	C0197417 (Biopsy cornea)	Exact
AAFA003	Brain	exeresis	C0006104 (Brain) and C0919588 (Exeresis)	Single to multiple
DGFA013	Aorta	laparotomie	C0003483 (Aorte)	Partial corres-pondance

Results

Using this approach, there were 5,212 (65%) CCAM codes out of the 7,926 CCAM codes used in this study that provided possible matching between the CCAM and the French terms in the UMLS. The results of each type of matching are displayed in Table 2.

There were 2,210 (27.5%) matches regarding the anatomical and action axes. On the other hand, there were 1,716 (21%) matches regarding only anatomical and 1,286 (16%) matches regarding only the action axis. Overall, 65% of the matching "anatomical terms" in the CCAM codes were matched to at least one UMLS Concept and 37% of the matching "action terms" in the CCAM codes were matched to at least one UMLS Concept.

Table 2- Results of each matching type

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

For the set of exact matching (n=200), 182 (91%) of matches between CCAM codes and UMLS concepts were rated as NT-BT and only in nine cases, the matches were rated as equivalent (see Table 3).

Table 3- Evaluation results of the "exact" set matchings type

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

For the set of single to multiple matchings (n=100), 61 and 44 of the anatomic and action axes respectively were equivalent to at least one UMLS concept. According to this type of matching, 27 (27%) matches between CCAM code and at

least one UMLS concept were rated as exactly equivalent, when 54 matchings were rated as NT-BT (see Table 4)

Table 4- Evaluation results of the "Single to Multiple" set matching type (n=100)

Single to multiple matching (100)	Equivalent	BT-NT	NT-BT	Incomplete	irrelevant
Anatomic	61	1	29	9	0
Action	44	0	49	1	6
Combinaison	27	0	54	10	9

Discussion

The CCAM is an important French terminology external to UMLS. The objective of this work is to partially map the CCAM to the UMLS. Because the CCAM terms are quite verbose (85% with strictly more than 4 terms), this task is difficult. Our approach using French NLP tools allows mapping 65% of the CCAM. In most of the cases, the qualitative evaluation has shown a NT-BT (narrower than) relation between a CCAM term and an UMLS concept. This result is easily explainable because terms the anatomic and action CCAM axes, which are mapped to the UMLS Metathesaurus, are generally broad terms (e.g. cornea for anatomy and resection for action).

Some fine-tuning of the method is possible. The use of the existing manual mapping between CCAM and MeSH performed by one author of this paper [21] can help find some matches with the UMLS. The impacts of the matching between UMLS and CCAM are: (a) possible matching with other terminologies (e.g. ICD10 used in French DRGs with CCAM); (b) querying PubMed from a CCAM code with a MeSH query (using CCAM-MeSH mapping)

Two main perspectives are identified: (a) the method presented here could be used with the MetaMap tool [22-23] in order to map the CCAM, and then compare our results with those obtained in [24];(b) to map CCAM to the "procedure" axis of the SNOMED International.

Acknowledgments

This work was partially supported through a grant by the InterSTIS project, funded by the French National Research Agency (ANR-07-TECSAN-010).

References

[1] Darmoni SJ, Joubert M, Dahamna B; Delahousse J, Fieschi M. SMTS® :a French Health Multi-terminology Server. AMIA symp, 2009; In press.

[2] Wang Y, Patrick J, Miller Ge, O'Hallaran J: A computational linguistics motivated mapping of ICPC-2 PLUS to SNOMED CT. BMC Medical Informatics & Decision Making 2008, 8(Suppl 1)**:** S5

[3] Rocha RA, Rocha BH, Huff SM. Automated translation between medical vocabularies using a frame-based interlingua. Proc 15th Annu Symp Comput Appl Med Care 1993:690-4

[4] Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. MD Comput 1990;7:104-9

[5] Fung KW, Bodenreider O. Utilizing UMLS for semantic mapping between terminologies. AMIA Annu Symp Proc 2005:266-70

[6] Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond Synonymy: exploiting the UMLS semantics in mapping vocabularies. Proceedings / AMIA Annual Symposium 1998:815-9

[7] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993; 32(4): 281-91

[8] Manchikanti L. CPT 2000: Interventional pain management coding in the new millennium. Pain Physician 2000; 3:73-85.

[9] The Nordic Medico Statistical Committee. NOMESCO Classification of Surgical Procedures. Uppsala and Copenhagen (2000).

[10] Trombert-Paviot B, Rodrigues J-M, Rogers J, Baud R, van der Haring E, Rassinoux A-M, Abrial V, Clavel L, Idir H. GALEN: a third-generation terminology tool to support a multipurpose national coding system for surgical procedures. International Journal of Medical Informatics 2000;58–9:71–85

[11] Rector AL, Baud R, Ceusters W, Claassen W, Rodrigues J-M, Rogers J, Rossi Mori A, van der Haring E, Solomon WD, Zanstra P. A comprehensive approach to developing and integrating multilingual classifications: GALEN's classification workbench. Journal of the AmericanMedical Informatics Association, 1998. Fall Symposium Special Issue: pp. 1115.

[12] CEN AFNOR EN NF 1828:2002. Health informatics - Categorial Structure for classifications and coding systems of surgical procedures.

[13] Rodrigues JM, Rector AL, Zanstra P, Baud R, Innes K, Rogers J, Rassinoux AM, Schulz S, Trombert Paviot B, ten Napel H, Clavel L, van der Haring E, Mateus C.An Ontology driven collaborative development for biomedical terminologies: from the French CCAM to the Australian ICHI coding system. Stud Health Technol Inform.2006;124:863-8.

[14] Zaiss A, Hanser S. The French Common Classification of Procedures CCAM. An option for Germany. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2007 Jul;50(7):944-52.

[15] Huff H. Exploring methods for improving the integration of LOINC in the UMLS. Report, University of Utah, Salt Lake City, USA 2003.

[16] Huff SM, Rocha RA, McDonald CJ, De Moor GJ, Fiers T, Bidgood WD, Jr., et al. Development of the Logical Observation Identifiers Names and Codes (LOINC) vocabulary. J Am Med Inform Assoc 1998;5(3):276-92

[17] Pereira S. Multi-terminology indexing of concepts in health.. PhD Thesis, University of Rouen, France 2008

[18] Soualmia LF. Study and evaluation of multiple approaches to expand queries for Information retrieval in Medecine. PhD Thesis, University of Rouen, France 2004.

[19] ISO 5964-1985. Documentation - Guidelines for the establishment and development of multilingual thesauri, International Organization for Standardization, Ref. No. ISO5964-1985

[20] World Wide Web Consortium, Simple Knowledge Organization System, http://www.w3.org/2004/02/skos/

[21] Massari P; Pereira S; Thirion B; Derville A & Darmoni SJ. Use of super-concepts to customize electronic medical records data display. Studies in Health Technology and Informatics, Volume 136, Pages 845 - 850, 2008

[22] Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesh TC, Wilbur WJ. The NLM Indexing Initiative. Proc AMIA Symp. 2000: 17-21

[23] Aronson AR. Effective mapping of biomedical text to UMLS Metathesarus : The MetaMap program. Proc AMIA Symp. 2001: 17-21

[24] Bousquet C; Sadou E; Merabti T; Trombert B; Kumar A;Darmoni S &Rodrigues JM.Mapping the French CCAM for Clinical Procedures to the UMLS metathesaurus using Galen ontology. Submitted to Medinfo2010.

Address for correspondence

Tayeb Merabti
CISMeF team, Rouen University Hospital
1, rue de Germont – 76031 Rouen, FRANCE
Cour Leschevin, Porte 21
E-mail: tayeb.merabti@chu-rouen.fr