Mapping the ATC classification to the UMLS Metathesaurus: some pragmatic applications

Tayeb Merabti, PhD[a,1], Hocine Abdoune MS[b], Catherine Letord, Pharm. D[a], Saoussen Sakji, PhD[a], Michel Joubert, PhD[b], Stefan J. Darmoni, MD, PhD[a]

[a] *CISMeF, Rouen University Hospital, Rouen. France & GCSIS, TIBS, LITIS EA 4108, Biomedical Research Institute, Rouen. France*
[b] *LERTIM, EA 3283,Faculty of Medicine, Mediterranean University, Marseille, France*

Abstract. ATC classification is a WHO international classification used to classify drugs. The aim of this paper is to evaluate two lexical methods in English and in French to map ATC to UMLS. Several implemented applications to illustrate the use of the ATC mapping in English and French: (a) MeSH translation in Norwegian, (b) Drug Information Portal, and (c) ATC to PubMed tool. Methods: Two lexical methods were used to map ATC to UMLS. The first approach used a French natural language processing tool to map French terms of ATC to the French terminologies of UMLS. The second approach used the MetaMap tool to map English terms of ATC to UMLS. The English MetaMap provides slightly more mappings than the French NLP tool (3,170 vs. 2,992). On the other hand, the French NLP tool provides a slightly better precision than MetaMap (88% vs. 86%). Using a manual mapping between ATC and MeSH. the union of the validated mappings between ATC and MeSH provides 2,824 mappings (68.7% of ATC codes of the fifth level). Lexical methods are powerful methods to map health terminologies to the UMLS Metathesaurus. Manual mapping is still necessary to complete the mapping.

Keywords. Abstracting Classification; Drugs; Internet; Multilingualism; Semantics; Terminology as topic; Unified Medical Language System; Vocabulary, controlled.

Introduction

The ATC (Anatomical, Therapeutic and Chemical) classification is an international classification[2] [1] used to classify drugs. The ATC classification is developed and maintained by the Collaborating Centre for Drug Statistics Methodology. Since 1982, the Centre is situated in Oslo at the Norwegian Institute of Public Health. The Centre is funded by the Norwegian government. In 1981, the WHO Regional Office for Europe

[1] Corresponding author: Tayeb Merabti, CISMeF, Rouen University Hospital, Cour Leschevin, Porte 21, 3ème étage, 1 rue de Germont, F76031 Rouen Cedex, France ; Email : tayeb.merabti@chu-rouen.fr.
[2] WHO Collaborating Centre for Drug Statistics Methodology: http://www.whocc.no/atcddd/

recommended the ATC system for international drug utilization studies. In 1996, WHO recognized the need to promote the use of the ATC system as an international standard for drug utilization studies. The Centre was therefore linked directly to the WHO Headquarters in Geneva instead of the WHO Regional Office for Europe in Copenhagen. The purpose of the ATC is to serve as a tool for drug utilization research in order to improve the quality of drug use. One component of this is the presentation and comparison of drug consumption statistics at international and other levels. The ATC classification is available in the following languages: English, French, German, Norwegian and Spanish.

The EU-FP7 project Patient Safety through Intelligent Procedures (PSIP)[3][2], runs from 2008 till 2011 and develops - among others - a prototype that should provide contextualized information and alerts as part of an electronic prescribing process in a hospital. For that purpose, the ATC classification was chosen to classify the drugs as it largely used in Europe.

The purpose of the Unified Medical Language System® (UMLS) [3] is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health. The main component of UMLS is the Metathesaurus which contains in English around 130 medical terminologies and proposes mappings between terminologies. Currently, although ATC is a WHO classification, it is not (yet) included in the UMLS Metathesaurus.

The process of terminology mapping consists of identifying identical (or approximately identical) concepts or relationships between terminologies [4],[5]. The objective of this work is to propose a mapping of the ATC classification to the UMLS Metathesaurus, using two tools in two languages: (1) using the MeTaMap lexical tools [5] in the English Language for the ATC classification; (2) using French lexical tools in the French language [6],[8] developed by the CISMeF team. The mapping that will be evaluated in this work will be the mapping between ATC and MeSH, which provides three pragmatic applications: (a) MeSH translation in Norwegian; (b) PSIP Drug Information Portal; (c) a software to access PubMed via an ATC code or a multi-lingual label.

1. Material and Methods

1.1. ATC classification

The WHO Collaborating Centre for Drug Statistics Methodology publishes a new issue of the complete ATC index annually. In ATC classification system, the drugs are divided into different groups according to the organ or system on which they act and their chemical, pharmacological and therapeutic properties. The ATC Code has the general form LCCLLCC where (L represents a letter and C a number). In this system, the drugs are classified in groups at five different levels: the 1st level: anatomical group (1 alphabetical character): fourteen main groups. The 2nd level: principal pharmacological/therapeutic group (2 numerical characters). The 3rd level: therapeutic/pharmacological sub-group (1 alphabetical character). The 4th level: chemical/therapeutic/pharmacological sub-group (1 alphabetical character). The 5th

[3] http://www.psip-project.eu/

level: sub-group for chemical substance: the individual active ingredient or the association of active ingredients (2 numerical characters).

For example:

N The nervous system
N05 Psycholeptics
N05B Anxiolytics
N05BA Benzodiazepine derivatives
N05BA01 Diazepam

The 2nd, 3rd and 4th levels are often used to identify pharmacological subgroups when that is considered more appropriate than therapeutic or chemical subgroups. Each level of this classification corresponds to an ATC code and an ATC label. The label of the 5th level corresponds to the International Generic Name of the substance, when it exists. International non-proprietary names (INN) are preferred. If INN names are not assigned, USAN (United States Adopted Name) or BAN (British Approved Name) names are usually chosen. Each code is allocated according to its principal indication. However, the latter can vary from one country to another, which explains why there may be several ATC codes for the same drug according to the concerned country. The main focus of the Drug Utilization Research Group in Europe was initially to improve drug utilization through cross-national drug utilization studies based on the ATC methodology [9]. Approximately 10% of the drugs do not have the same ATC code between France and Denmark (according to an internal study carried out by the VIDAL[4] company for the PSIP project). It was thus necessary to adapt to the French context and the Danish context to overcome with the problem of the "variable" ATC. This adaptation was made possible thanks to the participation of the Vidal company which provided the appropriate files; one correspondence table INN-ATC by country.

During the mapping module, the correspondence between the ATC classification and the MeSH terms (descriptors and Supplementary Concepts (SC)), was realized in order to find the best matching. The precision was 90% and the recall was 87%. Concerning the three different methods to automatically index ATC (method by title, method by brand name and method by indexation), 3,634 out of 5,073 of MeSH manually indexed resources and 1,341 out of 5,177 of MeSH automatically indexed resources were ATC automatically indexed. Most of the DIP resources are ATC automatically indexed by the method by brand names (51.4% for manually indexed resources and 24.4% for automatically indexed resources), followed by the method by title and the method by indexation.

1.2. Mapping to the UMLS Metathesaurus in English

MMTX is an implementation in the Java programming language of the MetaMap software [5] used to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text. Mappings to UMLS are associated to a score describing the similarity between terms to be mapped and the equivalent concepts in UMLS. For example, the ATC code "L01CA02 - Vinblastine" is mapped to the MeSH and SNOMED international terms "Vinblastine".

[4] http://www.vidal.fr/

1.3. Mapping to the UMLS Metathesaurus in French

French natural language processing tools and mapping algorithms were developed by the CISMeF team to map French health terminologies. These tools were used in previous works [7] and extended to link terms in multiple French health terminologies. This approach allows from a given term, to find a UMLS concept with French (or English) terms that are most lexically similar to it. Thus, to overcome some problems like inflections, stop-words, etc, basic natural language processing is necessary beforehand: (a) Removing stop words: frequent short words that don't affect the phrases such as "a", "Nos", "of", etc are removed from all terms; (b) 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, as compared to the stemming tools developed by the CISMeF team and the stemming tools in [7]; (c) The mapping used by this approach may provide three types of correspondences between all terms in source terminologies and French terms of the UMLS Metathesaurus: exact correspondence, single to multiple correspondence, partial correspondence. In this work, only the exact correspondence was evaluated. This relation may also to be represented into SKOS[5] (Simple Knowledge Organization System) language. SKOS language is also used to represent French health terminologies into the French Health Multi-terminological Server [10] to integrate the main health terminologies available in French, including those not yet mapped to the UMLS (e.g. ATC, CCAM, which is a French classification for procedures, ORPHANET, which is a thesaurus for rare diseases).

1.3.1. Exact correspondence

One ATC term and one French term in UMLS are in an "exact correspondence" if all words composing the two terms are exactly the same. Thus, according to this correspondence there is at most one UMLS Concept corresponding to the ATC term. Formally, an "exact correspondence" between this ATC term and one French term in UMLS is defined if the two terms are lexically similar. For example, the ATC code "A11HA06 - pyridoxal phosphate" is mapped to the MeSH and SNOMED international terms "pyridoxal phosphate".

2. Evaluation

We use an existing manual mapping between ATC codes of the 5th level and MeSH terms to evaluate the two approaches. Out of 4,268 ATC codes of the 5th level, a number of 4,108 (96%) mappings were performed manually by a librarian & pharmacist (CL) between ATC and MeSH terms. For each approach, three sets of codes were created: (1) the validated mappings: all the mappings of this set are obtained by the manual and the automatic approach; (2) the second set corresponds to all the mappings obtained automatically and not manually; (3) the last corresponds to all valid mappings not obtained automatically. Nevertheless, for the 4,108 manual mappings only 2,971 correspond to 1 to 1 mapping (one ATC code to one MeSH term). For example, the ATC code "J01MB06 - cinoxacin" is manually mapped to the MeSH

[5] http://www.w3.org/2004/02/skos/

term “cinoxacin”. Thus, in this evaluation we use only this number of mapping to evaluate the two approaches.

3. Results

3.1. Mapping to the UMLS Metathesaurus in English

Using this approach, there were 3,170 (74%) ATC codes out of the 4,268 ATC codes of the 5th level used in this study that are in an “exact matching” relation with at least one UMLS concept. Limiting this mapping to only the French terminologies included into UMLS, there are 3,062 (71%) ATC codes in an “exact matching” relations with at least one UMLS concepts. These codes are in “exact matching” with 3,062 MeSH preferred terms and 1,631 SNOMED International preferred terms.

3.2. Mapping to the UMLS Metathesaurus in French

Using this approach, there were 2,992 (70%) ATC codes out of the 4,268 ATC codes of the 5th level used in this study that are in an “exact matching” relation with at least one UMLS concepts. These codes are in “exact matching” with 2,499 MeSH preferred terms and 1.728 SNOMED International preferred terms.

3.3. Results of the evaluation

From the 3,170 mappings using English: (a) 2,740 (86%) mappings are in the set of manual mapping (validated mappings); (b) 430 (13%) mappings obtained automatically and not manually. Furthermore, 231 mappings were manually obtained and not automatically (see Table 1).
From the 2,992 mappings using French: (a) 2,640 (88%) mappings are in the set of manual mapping (validated mappings); (b) 352 (11%) mappings obtained automatically and not manually. Furthermore, 331 mappings were manually obtained and not automatically (see Table 1)..
Finally, the union of the validated mappings in French and in English provides 2,824 ATC to MeSH mappings (2,556 common in English and in French, 184 only in English, and 84 only in French) representing 68.7% of the 4,108 ATC codes of the fifth level. The use of the French lexical tool allows us to detect three misspellings in French ATC labels.

Table 1. Number of validated mappings and number of mappings obtained only automatically according to the English and French approaches.

Type of mappings	Number of validated mappings	Number of mappings obtained only automatically
English-based mapping	2,740 (86%)	430 (13%)
French-based mapping	2,640 (88%)	352 (11%)

4. Some practical applications

Already, three practical applications are under way using this ATC to UMLS mapping:

4.1. MeSH translation in Norwegian

The Norwegian Knowledge Centre for the Health Service[6] gathers and disseminates evidence about the effect and quality of methods and interventions within all parts of the health services. The Centre has decided to translate the MeSH to their language. Instead of starting from scratch, they will: (a) first, use the Swedish translation performed by the Karolinska Institute[7], as these two Nordic languages are quite similar; (b) second, they will use the proposed mapping ATC to MeSH as the ATC exists in Norwegian.

4.2. PSIP Drug Information Portal

The PSIP Drug Information Portal (DIP) was developed to allow French health professionals and patients to acccess relevant French information about drugs from main institutional health sites (e.g. French Drug Agency or European Drug Agency) [11]. The main innovation of this PSIP DIP relies on its multi-terminology indexing, mainly MeSH and ATC, but also using different codes, such as Chemical Abstracts Service (CAS) codes and its multi-terminology information retrieval based on the same terminologies and classifications.

4.3. ATC to PubMed

The third application is still under development. The objective is to create software to access PubMed via any ATC code (in any language). To do so, we have finalized the automatic mapping between ATC and UMLS by a manual mapping for the levels 1, 2, 3, 4 and 5 of the ATC. This manual mapping has been realized by a CISMeF medical librarian (CL) (N=5,359 (97%)). In most of the cases, this manual mapping was a 1 to N mapping (e.g. for the ATC code "D11AX18 - diclofenac", the MeSH mapping was "Diclofenac" and "Dermatologic agents"). For each ATC code, a predefined query was then created and could be launched on PubMed or on the PSIP DIP. For example, the PubMed query ""desensitization, immunologic"[MH] AND "allergens"[MH]" is associated to the ATC code "V01A".

5. Discussion

To the best of our knowledge, this is the first study to map ATC to UMLS using English (MetaMap) and French NLP tools. The English MetaMap provides slightly more mappings than the French NLP tool (3,170 vs. 2,992). The French NLP tool provides a slightly better precision than MetaMap (88% vs. 86%).

A number of algorithms and approaches have been proposed to create an automatic mapping between health terminologies [12]-[15]. For example, Rocha et al. [12] and Cimino et al. [13] both proposed a frame-based approach to perform mappings between health terminologies. Other approaches were proposed using UMLS [14] as a knowledge resource to perform mappings between terminologies making use of

[6] http://www.kunnskapssenteret.no/
[7] http://mesh.kib.ki.se/swemesh/swemesh_se.cfm

synonymy, explicit mapping relations and hierarchical relationships [15]. However, approaches using UMLS are limited to the biomedical terminologies already incorporated into UMLS.

Besides the ATC classification heavily used in Europe, RxNorm[8] [16] is a standardized nomenclature for clinical drugs and drug delivery devices. RxNorm is produced by the US National Library of Medicine. RxNorm provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software. RxNorm's standard names for clinical drugs and drug delivery devices are connected to the varying names of drugs present in many different controlled vocabularies within the UMLS Metathesaurus, including those in commercially available drug information sources. These connections are intended to facilitate interoperability among the computerized systems that record or process data dealing with clinical drugs. A mapping between ATC and UMLS imply an indirect mapping between ATC and RxNorm that should have some interest in the future. A formal evaluation of this mapping is mandatory before clinical use. This lexical methods were recently applied to one other French terminology (CCAM) not yet included in the UMLS Metathesaurus [17].

5.1. Limits

The current method is appropriate for ATC labels at the fifth level, which corresponds to its deepest level. This lexical tool was not adapted to the ATC other levels (level 1 to 4) because the ATC label used in the drug context, it is very difficult to process by NLP tool: e.g. the ATC code "digestive system" means in fact "drug therapy of digestive system diseases". Furthermore, in some cases, an ATC label of level 2, 3, 4 or 5 should also take into account the ATC label of level N-1, N-2 or N-3: e.g. for A12BA01 - Potassium chloride, must be distinguished from B05XA01 - Potassium chloride. A12BA01 is indicated in case of hypokalemia (by oral administration), therefore the A axis is taken into account (alimentary tract and metabolism). The final MeSH query for the A12BA01 - Potassium chloride is then: Potassium chlorside[MeSH] and hypokalemia/therapy[MeSH] and administration, Oral[MeSH].

Therefore, the mappings between ATC and MeSH for the ATC labels of level 1 to 4 were performed manually by the CISMeF pharmacist expert (CL). The overall mapping was then sent to the Collaborating Centre for Drug Statistics Methodology in Norway for validation.

The mapping between ATC and other terminologies is not so easy because: (a) mainly, one substance could have various ATC codes depending on whether it is used alone or in association, the diseases to be treated, the route of administration, (b) chemical classification varies from one terminology to another (e.g. mecamylamine is considered as an amine in ATC and as a terpenes in the MeSH); (c) the ATC classification is not purely anatomical (e.g. H axis stands for systemic hormonal preparations, excl. sex hormones and insulins); (d) in some cases, the MeSH lacks precision (e.g. impossibility to differentiate from beta blocking agents non-selective and selective; (e) in some cases, the MeSH hierarchy has to be carefully checked: e.g. Neomycin has three narrower terms: Framycetin, Paromomycin, Ribostamycin but there is also ATC codes for Neomycin and Framycetin. Therefore, the MeSH query for neomycin has to be

[8] http://www.nlm.nih.gov/research/umls/rxnorm/overview.html

restricted to Non exploded; (f) in some cases, pharmaceutical actions in the MeSH are not complete (e.g. the MeSH Supplementary Concept benfluorex has the following pharmaceutical actions: appetite depressants and antilipemic agents. The ATC classification provide another pharmaceutical action: hypoglycemic agents. Finally, some ATC codes are not mapped to the MeSH because there is no equivalent in that thesaurus (e.g. D11AC09 xenysalate).

6. Conclusion

Mapping the ATC classification to the UMLS Metathesaurus was performed with good results with automatic NLP and mapping tools. The coordinated use of appropriate NLP and semantic tools, international standards and ontology driven tools increased the quality of the mapping.

Acknowledgments

This work was partially granted by the PSIP project (Patient Safety through Intelligent Procedures in medication, 7th Framework Program of the European Union, Grant agreement n° 216130).

References

[1] Skrbo A, Begović B, Skrbo S. Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes. *Med Arh* **58** (2005), 138-41.

[2] Beuscart R, Hackl W, Nohr C. Detection and Prevention of Adverse Drug Events - Information Technologies and Human Factors, Amsterdam, 2009.

[3] Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. *Methods Inf Med* **32** (1993), 281–91.

[4] Bodenreider O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Res* **32** (2004), 267–270.

[5] Wang Y, Patrick J, Miller G, O'Hallaran J. A computational linguistics motivated mapping of ICPC-2 PLUS to SNOMED CT. *BMC Med Inform Decis Mak* **8 Suppl 1** (2008), S5.

[6] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Proc AMIA Symp* 2007, 17–21.

[7] Pereira S. Multi-terminology indexing of concepts in health. [Indexation multiterminologique de concepts en santé]. PhD Thesis, University of Rouen, France 2008.

[8] Merabti T. Methods for mapping medical terminologies: contribution to semantic interoperability between terminologies (Méthodes pour la mise en relation des terminologies médicales : contribution à l'interopérabilité sémantique Inter et Intra terminologique). PhD Thesis, University of Rouen, France, 2010.

[9] Bergman U. The history of the Drug Utilization Research Group in Europe. *Pharmacoepidemiol Drug Saf* **15** (2006), 95-8.

[10] Joubert, M; Dahamna, B; Delahousse, J; Fieschi, M & Darmoni SJ. SMTS® : Un Serveur Multi-Terminologies de Santé. Risques, *technologies de l'information pour les pratiques médicales : Informatique et santé* **17** (2009), 47-56.

[11] Sakji S, Letord C, Pereira S, Dahamna B, Joubert M, Darmoni J. Drug information portal in europe: information retrieval with multiple health terminologies. *Stud Health Technol Inform* **150** (2009), 497-501.

[12] Rocha RA, Rocha BH, Huff SM. Automated transaltion between medical vocabularies using a frame-based interlingua. *Proc Annu Symp Comput Appl Med Care* 1993, 690-4

[13] Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. *MD Comput* **7**(1990), 104-9

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

[15] Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond Synonymy: exploiting the UMLS semantics in mapping vocabularies. *AMIA Annu Symp Proc* 1998, 815-9.
[16] Parrish F, Do N, Bouhaddou O, Warnekar P. Implementation of RxNorm as a terminology mediation standard for exchanging pharmacy medication between federal agencies. *AMIA Annu Symp Proc* 2006:1057.
[17] Merabti T, Massari P, Joubert M, Sadou E, Lecroq T, Abdoune H, Rodrigues JM, Darmoni SJ. An Automated Approach to map a French terminology to UMLS. *MedInfo 2010, Stud Health Technol Inform* **160** (2010), 1040-4.