On the Annotation of Drug Databases with the Anatomical Therapeutic Chemical Classification System

Olivier Curé

On the Annotation of Drug Databases with the Anatomical Therapeutic Chemical Classification System

Olivier Curé

LIGM Université Paris-Est, France
{ocure}@univ-mlv.fr

Abstract. The Anatomical Therapeutic Chemical (ATC) classification system is frequently used to classify drugs according to an encoding system that considers the organ or system on which they act on and/or their therapeutic class and chemical characteristics. Motivated by the maintenance of a French drug database, we have transformed, using an inductive approach, the chemical layer of this classification into an OWL knowledge base. Using this knowledge base, we have been able to design interesting novel application functionalities. Nevertheless, these functionalities require that drugs are annotated in a different way than is generally done right now. In this paper, we highlight several French databases that are not fully benefiting from their ATC usage, present our annotation best practice and emphasize on the gains of its adoption.

1 Introduction

We have developed a self-medication Web application that provides information to the general public on mild clinical signs and associated Over The Counter (OTC) Drugs. This application is based on a drug database that stores almost all drugs sold in France. Being mainly used by end-users without some medical knowledge, the data quality of the database is a crucial aspect of our application, e.g., incorrect or missing contraindications on a drug description may impact the health of the patient. We have implemented a set of tools that help to check the data quality and assist the domain experts in a curation process when information are updated [1]. Our experience with pharmaceutical information motivated us to design these tools at the molecule level. That is, for each molecule of interest, we store its characteristics in terms of dimensions such as contraindications, side-effects, molecule interactions, warnings, therapeutic classes. To design such a knowledge base (KB), we have decided to adopt the Anatomical Therapeutic Chemical (ATC) classification system due to its wide adoption in drug databases. Note that several projects, *e.g.*, [4][3], propose to map this classification with other medical terminologies, *e.g.*, NDF-RT, and hence enable its integration in even more information sources. Being solely an encoding system that considers the organ or system on which they act on and/or their therapeutic class and chemical characteristics, none of the information needed for our data quality

system are natively proposed in ATC. Hence, we have decided to take advantage of the ATC encoding system to create a pratical pharmaceutical KB. This has been performed using a probabilistic inductive approach on the data stored on our drug database whose entities are annotated with identifiers of the ATC and other classifications, e.g. EphMRA. This approach is semi-automatic since it requires the supervision of a medical expert to validate the information attributed to a given molecule.

To keep our database up-to-date with the release, suppression and modification of drug products, we study the databases provided by several health organizations in France, e.g., Agence National de Sécurité du Médicament et des produits de santé (ANSM), and some over french databases, such as Vidal and BCB (e.g., Banque Claude Bernard). After studying these databases and their evolution for several years, we have a good understanding of their annotation practice with the ATC system. We argue that the organization of the ATC system motivates a form of annotation that prevents users from using its full potential. This is mainly due to the large numbers of identifiers corresponding to molecule combinations. This drawback not only restricts each drug database provider to design novel functionalities but also precludes data integration with other systems, e.g., linked data.

This paper is organized as follows. In Section 2, we present the ATC classification system. Then we detail our probabilistic inductive approach to design an OWL KB from our drug database. In Section 4, we highlight some of the benefits of adopting our KB to reasoning about molecules and drugs containing them. Section 5, introduces our simple best pratice in annotating drug products with the ATC system. We conclude and provide some future work in Section 6.

2 ATC

The ATC system (http://www.whocc.no/atcddd/) is an international classification of drugs and is part of WHO's initiatives to achieve universal access to needed drugs and their rational use. This classification takes the form of a tree structure with five different levels. The first level of the code is represented with a letter corresponding to one of the 14 predefined anatomical groups. The second level corresponds to a pharmacological/therapeutic subgroup. The third and fourth levels are chemical/pharmacological/therapeutic subgroups. Finally, the fifth level corresponds to chemical substances and supports the classification of drugs according to Recommended International Non-proprietary Names (rINN). We now provide an extract from the ATC hierarchy for some cough suppressants:

```
R: Respiratory system
 R05: Cough and cold preparations
   R05D: Cough suppressants, excluding combinations with expectorants
     R05DA: Opium alkaloids and derivatives
       ...
       R05DA08 Pholcodin
       R05DA09 Dextromethorphan
```

```
...
R05DA20 Combinations
```

The current version of ATC classification contains 5,658 identifiers out of which 4,406 are leaves and are hence supposed to identify molecules. Among these leaves, 16,3% do not identify a single molecule but rather describe a combination of molecules, e.g. R05DA20 in our previous example. Of course, annotating a drug with a 'combination' identifier does not provide much information on the active principle composing it.

3 Inductive creation of the ATC ontology

In this section, we detail our probabilistic inductive approach to transform the ATC classification into an OWL KB. This transformation has been performed using the DBOM system [2], a tool we have designed to exchange data between a relational database and a Semantic Web compliant KB. The result is an OWL DL ontology where each ATC identifier is transformed into an OWL concept, *owl:disjointWith* properties are declared mutually between all sibling concepts and the following properties are set between cascading concepts: *rdfs:subClassOf* for levels 5 and 4, 3 and 2, *hasGroup* for levels 4 and 3, *hasSystem* between levels 2 and 1.

In the following extract of our ATC ontology, we provide the description associated with the *Pholcodine* concept, with ATC code *R05DA08*. On line 1, we can see that this concept is identified by a given URI with a local name corresponding to *R05DA08* ATC code. Line 2 defines the concept associated with the *R05DA* code (*Opium alkaloids and derivatives*) to be a super concept of this concept. On lines 3 and 4, we present examples of *disjointWith* properties between sibling concepts, only the first and last concepts are displayed for brevity. Line 5 states a comment in the french language.

```
1. <owl:Class rdf:about="&p1;R05DA08">
2.  <rdfs:subClassOf rdf:resource="&p1;R05DA"/>
3.  <owl:disjointWith rdf:resource="&p1;R05DA01"/>
...
4.  <owl:disjointWith rdf:resource="&p1;R05DA20"/>
5.  <rdfs:comment xml:lang="fr">Pholcodine
6.    </rdfs:comment>
7. </owl:Class>
```

Starting from this OWL ontology, we can now perform an enrichment by inductive reasoning using our drug database. This database takes the form of a star schema where the central table stores information on general drug information, e.g., name, price, and is identified with a drug french identifier (CIP). This table is also related, using one-to-many or many-to-many associations of the Entity-Association logic model, to all kinds of medical information, e.g., side-effects, therapeutic classes, contraindications, including some ATC identifiers. The method aims to cluster relevant groups of products, generated using

the ATC hierarchy. Intuitively, we navigate in the hierarchy of concepts and create groups of products for each level, using the ATC to drug relation in our database. Then, for each group we study some specific domains which correspond to fields in SPCs (Summary of Product Characteristics), e.g. contraindications, and for each possible value in these domains we calculate the ratio of this value occurences on the total number of elements of the group. Table 1 proposes an extract of the results for the concepts of the respiratory system and the contraindication domain. This table highlights that our self medication database contains 56 antitussives (identified by code *R05D*), which are divided into 44 plain antitussives products (*R05DA*) and 12 antitussives in combinations (*R05DB*). For the contraindication identified by the number 76, i.e. allergy to one of the product's constituants, we can see that a ratio of 1 has been calculated for the group composed of the *R* ATC code. This means that all 152 products (100 %) of this group present this contraindication. We can also stress that for this same group, the breast-feeding contraindication (#9) has a ratio of 48 %, this means that only 72 products out the 152 of this group present this constraints.

We now consider this ratio as a confidence value for a given concept on the membership of a given domain's value. This membership is materialized in the ontology with the association of an concept to a property, e.g. the *hasContraIndication* property, that has the value of the given contraindication, e.g. breast-feeding (#9). In our approach, we only materialize memberships when the confidence values are superior to a predefined threshold θ, in the contraindication example we set θ to 60%.

This membership is only related to the highest concept in the concept hierarchy and inherited by its sub-concepts. For instance, the breast feeding contraindication (#9) is associated to the *R05* concept as its confidence value (83%) is the first column on line with *contraId* 9 that presents a θ superior to 60% in the *R* hierarchy. Also, the pregnancy contraindication (#21) is related to the *R05DB* ATC concept since its value is (73%).

Using this simple approach, we are able to enrich the ATC ontology with ax-

Table 1. Analysis of contraindications for the respiratory system

	R	R05	R05D	R05DA	R05DB	..	R05DA09	..
occurences	152	71	56	44	12	..	21	..
ContraId								
9	.48	.83	.86	.82	1		1	
21	.26	.39	.3	.2	.73		0	
76	1	1	1	1	1		1	
108	.34	.69	.84	.84	.82		1	
109	.35	.66	.8	.8	.82		1	
110	.34	.73	.89	.86	1		1	
112	.34	.71	.88	.86	.91		1	

ioms related to several fields of SPCs. At the end of this enrichment phase, the expressiveness of the newly generated ontology still corresponds to OWL DL.

This method can easily be applied to other drug related classifications, e.g. EphMRA, as soon as we consider that the ontology is presented in a DL formalism and a relation relates CIPs to identifiers of this ontology.

4 Features supported by the Knowledge Base

In this section, we highlight three important functionalities that have been implemented using our medical KB. The first two functionalities enable to improve the data quality of the drug database while the last one supports data integration to over KBs.

The first one consists in reasoning at the molecule level over updates performed on the drugs of our database. That is, whenever a drug composed of a molecule represented in our KB is inserted, updated or deleted in our database, the system automatically checks whether this modifications will produce a consistent database instance. This is performed using a semantic query expansion approach that produces SPARQL queries from inferences over the knowledge.

The second feature is related to the automatic enrichment of a drug SCP given the knowledge stored in our molecule KB. That is, our system administrator can now insert a drug name and its composition and the system automatically inserts properties associated to the molecules composing the drugs. We found out that even in the case of compound drugs, this approach is relevant and produces an important amount of the information expected from this drug's SPC. Note that all of these data updates are double-checked by medical experts for data quality reasons. Nevertheless, this semi-automatic approach enables us to improve the soundness and completeness as well as save a lot of time on the database maintenance issues.

Finally, the design of this KB together with the fine-grained annotations of drug databases supports a better data integration with other terminologies and KBs. For instance, we have submitted a project at ISWC 2012's Semantic Web Challenge (finishing at the 3rd place) on the implementation of a self-medication application using Linked Open Data data sets. In that Web application, we were checking the consistency of the proposed information based on a mapping between the Drugbank terminology and our molecule KB.

5 ATC annotation best practice

This section claims that organizations supporting the maintenance of a drug database, at least in France, do not take full benefits of annotating drug with the ATC. For instance, the french ANSM organization maintains drug databases containing 16,516 drug products out of which 13,358 contain a single molecule (i.e., we do not consider excipients here which are generally not represented in ATC, but that is another problem we are not considering in this paper). That is 3,159 drug products contain more than one molecule, representing 19,2% of the

total number products available in France. Most of these drugs are either not encoded with the ATC or annotated with a 'combination' identifier. Note that this proportion is also encountered in the drug databases maintained by Vidal and BCB, two important medical providers in France.

In order to create and maintain a KB as presented in Section 3 and implement application features as presented in Section 4, one has to annotate drug products with each molecule that is composing the drugs rather than a combination that does not describe sufficiently the product in question. For instance, the *R05DA20* code identifies compound chemical products combining *opium alkaloids* with other substances. The *Hexapneumine* syrup is generally annotated with this code but it is inaccurate since it does not quantify and qualify the contained molecules. We argue for another approach where this drug is classified by the conjunction of its compounds, i.e. *pholcodin*, *chlorphenamin* and *biclotymol* which respectively correspond to R05DA08, R06AB04 and R02AA20. In terms of a database modeling, this approach only requires to define one-to-many relationship between a drug and its ATC identifiers rather than defining a single column that stores an ATC code in the drug table.

6 Conclusion

In this paper, we emphasized that providing a fine-grained annotation of drug products with the ATC classification opens possibilities in developing interesting functionalities in medical applications. This approach mainly invites drug database producers to annotate with ATC identifiers corresponding to molecule identifiers rather than with 'combinations' ones. This form of annotation has been quite positive in our self-medication application and helped to improve the overall data quality and reduce the time our health experts spend on curating the database. As future work, we started to enrich our own molecule KB with chemical products that are found as excipients in drug compositions. The idea is to describe excipients that are known to have side-effects or contraindications, e.g. aspartam and phenylcetonuria. This extension to the molecule KB will help in improving our data quality approach.

References

1. O. Curé. Improving the data quality of drug databases using conditional dependencies and ontologies. *J. Data and Information Quality*, 4(1):3, 2012.
2. O. Curé and J.-D. Bensaid. Integration of relational databases into owl knowledge bases: demonstration of the dbom system. In *ICDE Workshops*, pages 230–233, 2008.
3. F. Mougin, A. Burgun, and O. Bodenreider. Comparing drug-class membership in atc and ndf-rt. In *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, pages 437–444, New York, NY, USA, 2012. ACM.
4. Q. Zhu, G. Jiang, L. Wang, and C. G. Chute. Standardized drug and pharmacological class network construction. In *MedInfo*, page 1125, 2013.