Ontology-based Data Quality enhancement for Drug Databases

Olivier Curé
Université de Marne-la-Vallée
5 bd Descartes 77454
Marne-la-Vallée, France
ocure@univ-mlv.fr

Jean-Paul Giroud
Pharmacology department
CHU Cochin
Paris, France
jeanpaul.giroud@cch.ap-hop-paris.fr

ABSTRACT

This paper proposes a solution to enhance data quality features in drug databases. The method we are using is based on the integration of controlled terminologies for the drug domain. This approach requires several steps : (1) transforming the terminologies into a Semantic Web compliant Description Logics, namely OWL DL, (2) associating new axioms to concepts of these ontologies based on inductive reasoning on drug databases, and (3) assisting, using these OWL ontologies, our team of health care professionals in detecting and repairing data quality violations of selected features in drug databases. This last step aggregates drug products using ontology concepts and checks the characteristics of the products with the concept's properties. We present a concrete example of this solution on a self-medication application using the European Pharmaceutical Market Research Association and Anatomical Therapeutical Chemical terminologies.

Keywords

Ontologies, Integration, Inductive reasoning, Data quality, Drugs, Self-medication

1. INTRODUCTION

The Internet already has a significant impact on the industrial countries health care systems; much of it determined by how the World Wide Web influences the patient-physician relationship [2]. It is currently known that patients on the web have very conservative values : trust, quality, reliability and loyalty. Considering the needs of those web patients must be a key factor for any health related web application. Replying to such expectations requires an in-depth collaboration between health care professionals and computer scientists.

Such a collaboration has allowed to develop a medical informatics web application for the domain of self-medication, i.e. the act for a patient to treat oneself, with or without drugs. This patient oriented application provides information and services on mild clinical signs and related medications. Among available services, the system proposes drug advices for the consumer considering a symptom description, detection of drug interactions, selection of cost-effective treatments and the ability for the patient to maintain a simplified health care record (SEHR). In order to be useful for the patient, these services have to exploit and to propose information that comply to the previously introduced patient requirements.

In this paper, we are concentrating on the quality aspect of patient's needs. For this purpose, we are being influenced by ISO 9000 Quality standards and consider two data quality elements : completeness and correctness. In terms of completeness, we are interested in the presence or absence of data in the dataset. We distinguish between two aspects of completeness : (i) comission, i.e. when excess data are present in the dataset, and (ii) omission, i.e. when some data are absent from the dataset. Considering the correctness aspect, we expect the dataset to contain the correct data. In the domain of drug information, we relate these quality elements to the dataset of the Summary of Product Characteristics (SPC). The SPC contains a description of a certain medicinal products properties and the conditions attached to its use. In this research, the SPC fields we are the most interested in are drug contra-indications, side-effects and cautions in using a given drug. For example, considering the contra-indication field of a given drug's SPC, we relate the quality elements in the following way:

- with comission, we do not want to have excess entries in the contra-indication list.
- with omission, we do not want some contra-indication entries to be absent from the list.
- with correctness, we expect the correct set of entries to be associated to this drug contra-indication's list.

The respect of these data quality elements is essential in a domain as sensible as self-medication where incorrect data may have important consequences on the patient's health.

In order to ensure these data quality features, we propose the integration of health care terminologies related to the drug domain. These terminologies are then used to reason other our drug database and to possibly detect incompleteness and incorrectness in the dataset. The terminologies we are integrating are the most frequently used in french drug data sources. They are usually intended for high-level indexing of medication information and as a coding solution. Thus it is interesting to compare our way of classifying drugs using these terminologies with other data sources. This approach also highlights some limitations in the use of these terminologies. In this paper, we are addressing some of these problems and propose some solutions based on relevant modeling approaches.

WWW2007, May 8–12, 2007, Banff, Canada.
.

In order to study the data quality of our drug database, the data contained in these terminologies is not sufficient. We propose a way to enrich these terminologies using inductive reasoning on drug databases. With this solution, we transform the terminologies into ontologies by providing some formal presentation of definitional drug information.

This paper is organized as follows. In Section 2, we motivate our approach in the context of our self-medication application and the drug information domain. In Section 3, we present our method to enrich drug related terminologies, so that we can consider them as ontologies, using inductive reasoning on our drug database. In Section 4, we introduce our solution to detect inconsistencies in the drug database using our two previously enriched drug terminologies. In Section 5, we propose some improvements that enhance the detection and repairing of data quality violations. We conclude with Section 6 and propose some future works on the usage of the drug ontologies.

2. MOTIVATING EXAMPLE

In this section, we motivate our approach toward data quality assessment in a self-medication application and its drug database. We present the main features of this application and stress the need for reasoning facilities. Finally, we introduce the main characteristics of the drug terminologies exploited in our solution to data quality improvement.

2.1 The self-medication application

The origin of the self-medication project lies in Pr. Jean-Paul Giroud and Dr. Charles Hagège's experience in self-medication and pharmocology. Over the last two decades, they released several best-selling books in this domain. The last edition of the guide [12] covers mild symptoms and associated medication on the first part while the second part details information on all drugs available on the french market, i.e. OTC (Over The Counter) and non-OTC products, allotherapy, homeopathy, phytotherapy, etc. Two types of searching the book can be observed : (1) searching for a drug product and studying its information, (2) searching for a symptom and looking for available medications.

The success of the books motivated a collaboration, between the teams of the department of clinical pharmacology at the Cochin hospital in Paris, headed by Pr. Giroud, and computer scientists from the University of Marne-la-Vallée. The objective of this collaboration is to develop a web application supporting self-medication. The purpose of this application is to support a safe practice of self-medication and to educate the patient on mild clinical signs and drug products. Thus another objective of this approach is to enhance the communication between general practitioners and patients by educating the general public.

The central component of the application is a drug database which stores all the SPCs of drug products sold in France. The database also stores additional information for each drug, i.e. a drug rating based on an efficiency/tolerance ratio and an opinion written by our team of health care professionals. In terms of data exploited by this application, SPC is a first class citizen. A main problem about the data contained in SPCs is their content inadequacy toward an unambiguous patient understanding. This problem leads patients to a partial reading and understanding of SPCs. This aspect increases increases possibilities of health problems due to interactions with other drugs, contra-indications, etc. This problem has been addressed early in the Giroud-Hagège book series by translating idiosyncratic medical terms into easily understandable terms for the patient.

Figure 1 proposes an overall view of the self-medication application. A first aspect is the maintenance of the drug database, a task performed by our team of health care professionals. This task is burdensome because of the rapid evolution, i.e. weekly rate, of the drug market with emergence and withdrawal of drugs, modification of drug's indications and composition, drug switches, social security reimbursement rate changes, etc. In order to facilitate the work of our team of health care professionals, we designed a web interface that enables to efficiently update the drug database. The maintenance efficiency is a priority for us as our drug database is the central component of several medical applications.

Considering the self-medication application, it is necessary to transform this drug database into a Description Logics (DLs) compliant knowledge base (KB). DLs are a family of knowledge representation formalisms with well-understood semantics and computational properties [3]. The transformation to this formalism is required by the need to use a DL reasoner, namely Pellet [16], to perform some inferences in some self-medication services, e.g. decisions related to anamnesis, maintenance of the patient's SEHR [7]. This transformation task is performed by the DBOM system [6, 8]. In a nutshell, DBOM has been developed as a Protégé [11] plug-in and enables to create semi-automatically and to maintain automatically OWL KBs from relational databases. The main services proposed by the self-medication application are the maintenance of the SEHR and the 'diagnosis' service.

The overall goal of the SEHR is to store health related information concerning a particular patient. The SEHR is an XML document that stores three categories of patient information :

- general information concerning a patient such as name, gender, date of birth, etc..
- medical information mainly concerning known diseases, allergies, etc..
- drug consumption information which distinguishes consumption start date, dosage and duration of treatment. The system supports the description of discrete and continuous (e.g. life long treatments) prescriptions. Information concerning the source of prescription, either a health care professional or self-medication, are also provided to emphasize trustworthiness of the data stored.

Whenever a patient modifies her SEHR, several reasoning procedures, which require the self-medication KB, are performed to detect possible interactions between drug treatments or contra-indications with declared diseases or allergies.

The SEHR document is also exploited in the 'diagnosis' service. In this module, the patient interacts with the application to describe a symptom. The system then performs some inferences using the SEHR, the self-medication KB and the results of the patient Q&A to provide a list of efficient OTC drug products ordered by their ratings. The patient can then select a drug and study the complete SPC with

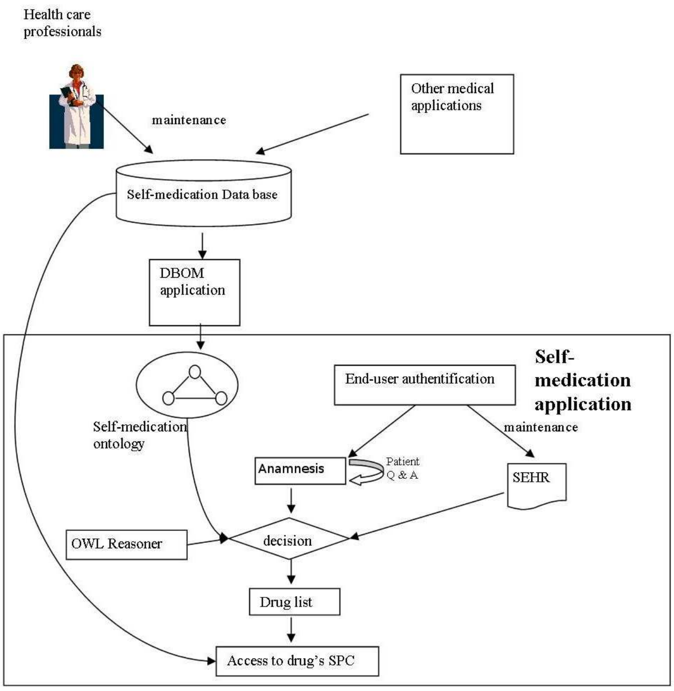


Figure 1: General architecture of the self-medication application

extra information such as opinion, price, social security reimbursement rate, etc. In any case a situation is detected to be out of the range of self-medication, the system asks the patient to consult a general practitioner and does not provide, for ethical reasons, further drug and symptom information.

2.2 Drug terminologies

Controlled terminologies and classifications are widely available for health care [5]. In this paper, we are interested in the drug related classifications that are the most frequently used in french drug databases. Namely these are the European Pharmaceutical Market Research Association (EphMRA) © and the Anatomical Therapeutic Chemical (ATC) © classifications. As most french drug databases, we use the CIP french code system to identify products. For instance, the drug *Tussidane*© syrup is sold as a 250ml bottle product with the CIP identification 3622682 and a 125ml bottle product which is identified by CIP 3622676. So each product has a distinct CIP code and many CIPs may be available for a given drug.

2.2.1 EphMRA classification

The EphMRA brings together European, research-based pharmaceutical companies operating on a global perspective. One of the missions of the EphMRA is to provide recognised standards by continuously supporting and actively participating in establishing high levels of standards and quality control in pharmaceutical marketing research. The Anatomical Classification system (AC-system) is the main classification developed by the EphMRA, with its sister organisation in the USA, the Pharmaceutical Business Intelligence and Research Group (PBIRG). This system represents a subjective method of grouping certain pharmaceutical products. The products are classified according to their main therapeutic indication and each product is assigned to one category. In the AC-system, categories are organized on a cascade of 4 levels where each sub-level gives additional details about its upper-level.

The first level of the code is based on a letter for the anatomical group and defines 14 groups, e.g. A for Alimentary tract and metabolism, B for Blood and blood forming organs, etc.

The second level is used to regroup several classes together, in order to classify according to indication, therapeutic substance group and anatomical system. This level adds a digit to the letter of the first level and enables the creation of the cascade classification. Therefore, before creating a new second level, all existing possibilities of classification should be analyzed. There could be cases where it is necessary to create a second level without a cascade to the third or fourth level. However, these cases are seldom in the current classification.

The third level adds a letter to a second level code and describes a specific group of products within the second level. This specification can be a chemical structure or it can describe an indication or a method of action.

The fourth level gives more details about the elements of the third level (formulation, chemical description, mode of action, etc.). Fourth level codes add a digit to third level ones.

The complete hierarchy for antitussive drugs corresponds to :

```
R: Respiratory system
```

```
 R5: Cough and cold preparations
   R5D: Antitussives
      R5D1: Plain antitussives
      R5D2: Antitussives in combinations
```

2.2.2 ATC classification

The ATC system [17] proposes an international classification of drugs and is part of WHO's initiatives to achieve universal access to needed drugs and rational use of drugs. In this classification, drugs are classified in groups at five different levels. In fact, the ATC system modifies and extends the AC-system of EphMRA. Thus the first level is composed of the 14 groups of the EphMRA system. The second is also quite similar and corresponds to a pharmacological/therapeutic subgroup. The third and fourth levels are chemical/pharmacological/therapeutic subgroups. Finally, the fifth level corresponds to the chemical substances. With its fifth level, the ATC classification enables to classify drugs according to Recommended International Nonproprietary Names (rINN). This is different from EphMRA's classification where the leafs of the tree (fourth level) give details on a wider perspective (formulation chemical description, mode of action, etc.). Thus we consider that the use of both terminologies in our data quality assessment approach is complementary and relevant in our inductive approach.

We now provide an extract from the hierarchy of cough suppressants in ATC system:

```
R: Respiratory system
 R5: Cough and cold preparations
   R05D: Cough suppressants, excluding combinations
   with expectorants
     R05DA: Opium alkaloids and derivatives
       R05DA01 Ethylmorphine
       ...
       R05DA08 Pholcodine
       ...
       R05DA20 Combinations
```

The *R05DA20* code identifies compound chemical products that combine *opium alkaloids* with other substances. An example for this code is the *Hexapneumine*© syrup which contains the following chemical substance : *pholcodin*, *chlorphenamin* and *biclotymol* which respectively correspond to R05DA08, R06AB04 and R02AA20 ATC codes. In Section 5, we argue on the relevance of classifying products with these compound codes and propose a more satisfying solution.

3. TERMINOLOGY ENRICHMENT USING INDUCTIVE REASONING

The purpose of the terminology integration is to enable the pooling of products in a coherent way such that common properties can be inferred. Thus given a code of either of our two terminologies, we can aggregate products in a sufficiently coherent manner. The issue we are now facing is that these terminologies do not propose any valuable properties.

This section aims to provide drug related properties, e.g. contra-indications, side-effects and cautions, to these terminologies using inductive reasoning on the drug database. In order to perform such enrichment, it is first necessary to transform these classifications into a DL formalism. This step has been again performed using the DBOM system. The results are OWL DL ontologies [9] where:

- each EphMRA code (respectively ATC code) is transformed into an OWL concept
- *rdfs:subClassOf* properties are set between cascading concepts
- *owl:disjointWith* properties are declared mutually between all sibling concepts
- a *rdfs:comment* property is set in the french language for this concept.

In the following extract of our ATC ontology, we provide the description associated with the *Pholcodine* concept, identified by 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 reasons. 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 these OWL ontologies, we can now perform the enrichment by inductive reasoning. In Figure 2, we present an extract from a drug database instance which emphasizes on relations involved in the enrichment for the contra-indication and therapeutic class aspects. Other enrichment possibilities, e.g. side-effects and cautions, require other relations which are also accessible from our database.

Figure 2a proposes a subset (CIP and product name) of the columns available in the *Drug* relation where two drug products are displayed. Figure 2b presents an incomplete list of the *ContraIndication* relation which stores all terms related to drug contra-indications. Now the *ProductContraIndication* relation enables to relate products identified by CIPs with their contra-Indications (Figure 2c). The *TherapeuticClass* relation regroups all therapeutic classes encountered in the drug domain and identifies them with integer values (Figure 2d). This identifier is related to CIP codes in the *ProductTherapeutic* relation (Figure 2e). Finally, two relations relate EphMRA and ATC codes to CIPs, respectively *ProductEphMRA* (Figure 2f) and *ProductATC* (figure 2g) relations.

In the following, we present the inductive reasoning method on the EphMRA ontology, also named AC-ontology, and stress that an adaptation for the ATC ontology is obvious. The method used to enrich the AC-ontology is based on induction reasoning on relevant groups of products, generated using the AC-hierarchy. Intuitively, we navigate in the hierarchy of AC-concepts and create groups of products for each level, using the ProductEphMRA relation (Figure 2f). Then, for each group we study some specific domains which

Table 1: Analysis of contra-indications for the respiratory system

	R	R05	R05D	R05D1	R05D2
occurences	152	71	56	44	12
ContraId #					
9	.48	.83	.86	.82	1
21	.26	.39	.3	.2	.73
76	1	1	1	1	1
108	.34	.69	.84	.84	.82
109	.35	.66	.8	.8	.82
110	.34	.73	.89	.86	1
112	.34	.71	.88	.86	.91

correspond to fields in SPCs, e.g. contra-indications, 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 contra-indication domain. This table highlights that our self medication database contains 56 antitussives (identified by AC-code *R05D*), which are divided into 44 plain antitussives products (*R05D1*) and 12 antitussives in combinations (*R05D2*). For the contra-indication identified by the number 76, i.e. allergy to one of the constituants of the product, we can see that a ratio of 1 (100%) has been calculated for the group composed of the *R* AC code. This means that all 152 products of this group present this contra-indication. We can also stress that for this same group, the breast-feeding contra-indication (#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 AC-concept on the membership of a given domain's value. This membership is materialized in the ontology with the association of an AC-concept to a property, e.g. the hasContraIndication property, that has the value of the given contra-indication, 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 contra-indication example we set θ to 0.6 (60%).

This membership is only related to the highest concept in the AC hierarchy and inherited by its sub-concepts. For instance, the breast feeding contra-indication (#9) is associated to the *R05* AC-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 contra-indication (#21) is related to the *R05D2* AC concept since its value is 0.73 (73%).

Using this simple approach, we are able to enrich the AC-ontology with axioms related to several fields of SPCs. At the end of this enrichment phase, the expressiveness of the newly generated ontology still corresponds to an OWL DL ontology. The following code proposes an extract of the AC-ontology, in RDF/XML syntax, where we can see the definition of *R05D2* concept (line #1 to #12). This description states that the concept:

- has the contra-indication identified by *CI_21* (lines 2 to 7) which corresponds to pregnacy (lines 13 to 16).
- is a subconcept of the *R05D* concept (line 8)
- is disjointWith the concept identified by the *R05D1* code (line 9)
- has a comment, expressed in the french language (lines 10 and 11).

```
1. <owl:Class rdf:about="&j.0;R05D2">
2.  <rdfs:subClassOf>
3.   <owl:Restriction>
4.    <owl:onProperty
         rdf:resource="&j.0;hascontraIndication"/>
5.    <owl:hasValue rdf:resource="&j.0;CI_21"/>
6.   </owl:Restriction>
7.  </rdfs:subClassOf>
8.  <rdfs:subClassOf rdf:resource="&j.0;R05D"/>
9.    <owl:disjointWith rdf:resource="&p1;R05D1"/>
10.  <rdfs:comment
         xml:lang="fr">ANTITUSSIFS EN ASSOCIATION
11. </rdfs:comment>
12.</owl:Class>
13.<j.0:contraIndication rdf:about="&j.0;CI_21">
14. <rdfs:comment xml:lang="fr">grossesse
15. </rdfs:comment>
16.</j.0:contraIndication>
```

This method can easily be applied to the ATC ontology or other drug related ontologies as soon as we consider that the ontology is presented in a DL formalism and a relation relates CIPs to identifiers of this ontology.

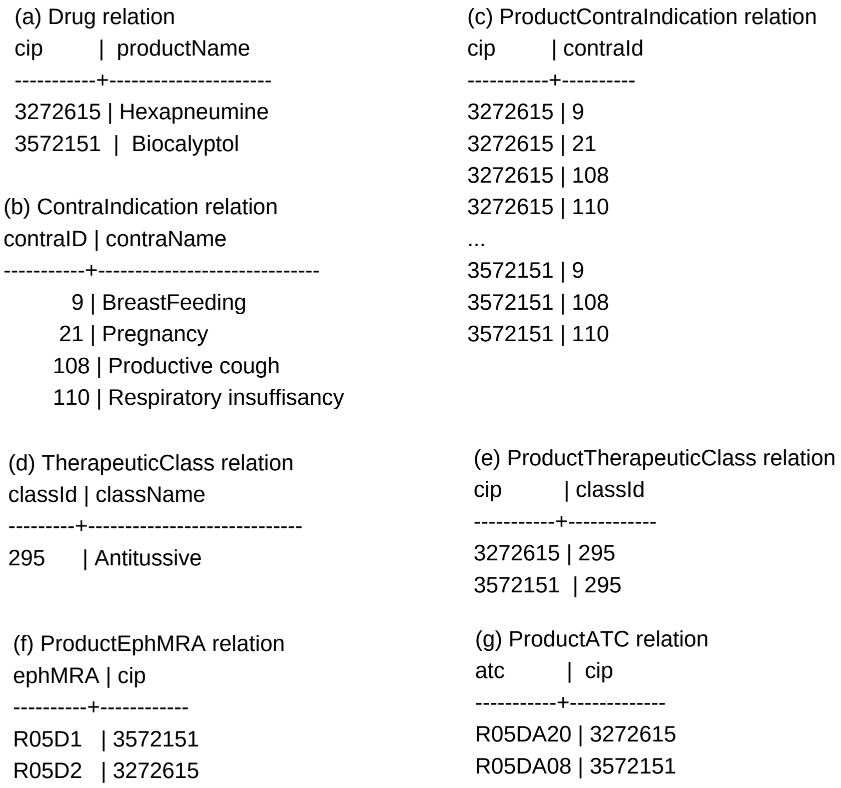
4. ONTOLOGY-BASED DETECTION AND REPAIRING

4.1 Detection method

In this section, we only consider data quality violations at the completeness (comission and omission) and correcteness levels. The principle we use to detect these violations are supported by the ontologies defined in Section 3 and the drug database, from which Figure 2 is an extract.

The main assumption of this method is the following. We consider that the drug database presents overall good data quality. This is the reason why we designed the ontology enrichment from induction on this database. But as the domain of providing information on self-medication drugs is so sensible, we want to improve the data quality of this given database. Thus we are using the properties associated to the concepts of our two ontologies to detect data quality violations. The potential of this approach is interesting because drugs can generally be aggregated using different characteristics, e.g. therapeutic class, chemical substances, etc. We believe that an efficient approach to design relevant groups are based on the use of EphMRA and ATC ontologies.

We can view the relation between the ontologies and the relational drug database with a logical point of view. The schema part of a DL KB is typically called a TBox (terminology box) and is a finite set of universally quantified implications [3]. Most DLs can be considered as decidable fragments of first-order logic. Thus their axioms have an equivalent representation as first-order-formulae [4]. On the other hand, the schema of a relational database is defined in terms of relations and dependencies [13], also named integrity constraints. In [1], the authors explain that most dependencies can be represented as first-order formulae and



```
(a) Drug relation
cip        |  productName
-----------+----------------------
3272615 | Hexapneumine
3572151  |  Biocalyptol
```

```
(b) ContraIndication relation
contraID | contraName
-----------+------------------------------
        9 | BreastFeeding
      21 | Pregnancy
    108 | Productive cough
    110 | Respiratory insuffisancy
```

```
(c) ProductContraIndication relation
cip        | contraId
-----------+----------
3272615 | 9
3272615 | 21
3272615 | 108
3272615 | 110
...
3572151 | 9
3572151 | 108
3572151 | 110
```

```
(d) TherapeuticClass relation
classId | className
---------+-----------------------------
295      | Antitussive
```

```
(e) ProductTherapeuticClass relation
cip         | classId
-----------+------------
3272615 | 295
3572151  | 295
```

```
(f) ProductEphMRA relation
ephMRA | cip
----------+------------
R05D1   | 3572151
R05D2   | 3272615
```

```
(g) ProductATC relation
atc          |  cip
-----------+-------------
R05DA20 | 3272615
R05DA08 | 3572151
```

Figure 2: Extract of the self-medication database

have a dual role in relational databases : they describe the possible worlds as well as the states of the databases. Reiter also observed in [15] that integrity constraints are sentences about the state of the database and are not objective sentences about the world. As discussed in [14], although the expressivity of DLs underlying OWL and of relational dependencies is clearly different, the schema languages of the two are quite closely related.

The interpretation of schemas in both DLs and relational databases are grounded in standard first-order semantics. In this semantics, a distinction is made between legal and illegal relational structures. A structure is legal when it satisfies all axioms defined in its schema, otherwise it is illegal. The terms used to denote legal structures in DLs and relational databases are different, respectively *models* and *database instances*.

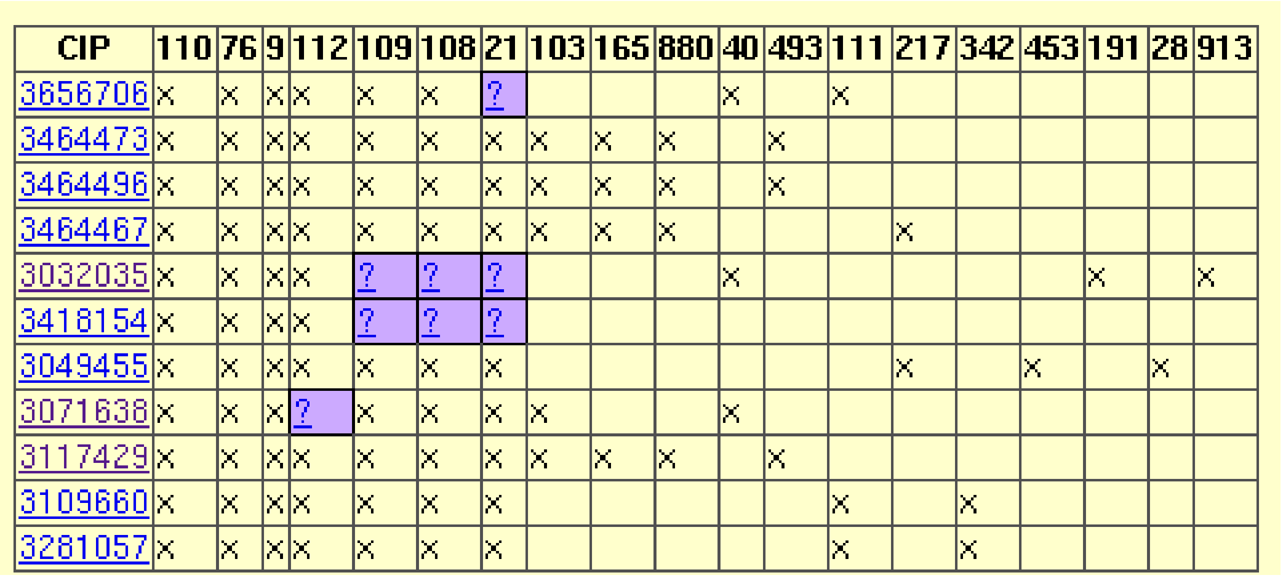
Whenever a relational database is updated, its dependencies are interpreted as check. If the check is satisfied, the database instance is modified accordingly otherwise the update is rejected. The behavior on the models of DLs first-order formulae is different [10, 14].

In our approach, we consider that each drug database instances respect a given set of integrity constraints. But we also require from these database instances to respect a set of dependencies expressed in some associated ontologies, EphMRA and ATC ontologies in this paper. Thus we want the drug database instances to satisfy explicitly provided database schema integrity constraints as well as the DL dependencies. The mechanism proposed for the latter has to deal with the context of the drug domain where many exception occurs. For example, in the case of processing a group of drugs based on a given EphMRA concept, we may encounter drugs which do not present the same set of contra-indications. This may be caused by the dosage of the drugs, an aspect we are currently trying to solve with the integration of the DDD system, or the presence of excipients, an issue we aim to address with the integration of rules in ontologies.

In the current solution, in order to deal with these exceptions, we must involve our team of health care professionals (HCPs) in the process of repairing violations. Thus our detect and repair approach is semi-automatic: detection is automatic and repairing involves the validation of the HCPs. In order to facilitate these tasks for our HCPs, we designed a web interface that highlights potential violations for a set of drugs and proposes a fast and easy way to repair them.

We are actually proposing two graphical solutions to repair these violations:

- a group-centric approach where a code in either of our two terminologies can be selected from a web frame. In a second web frame, matrices of CIPs and identifiers of SPC fields, e.g. contra-indications, are displayed for the selected terminology code. This frame also proposed the names corresponding to each identifiers of SPC fields. Figure 3 proposes an extract from this presentation for the contra-indication fields for the *R05D2* EphMRA code. From such a matrix, it is possible to click on a CIP number and to access the complete SPC of the drug. In this situation, the composition field may help the health care professional to take a deci-



CIP	110	76	9	112	109	108	21	103	165	880	40	493	111	217	342	453	191	28	913
3656706	x	x	x	x	x	x	?				x		x						
3464473	x	x	x	x	x	x	x	x	x	x		x							
3464496	x	x	x	x	x	x	x	x	x	x		x							
3464467	x	x	x	x	x	x	x	x	x	x				x					
3032035	x	x	x	x	?	?	?				x						x		x
3418154	x	x	x	x	?	?	?												
3049455	x	x	x	x	x	x	x							x		x		x	
3071638	x	x	x	?	x	x	x	x			x								
3117429	x	x	x	x	x	x	x	x	x	x		x							
3109660	x	x	x	x	x	x	x						x		x				
3281057	x	x	x	x	x	x	x						x		x				

Figure 3: Extract from a group-centric view :contra-indication matrix for the *R05D2* EphMRA concept

sion. Also in Figure 3, interrogation marks highlight possible violations. For example, on Figure 3, two violations are possible for the contra-indication, identified with value 108 (*productive cough*) : for products identified by CIPs 3032035 and 3418154. This detection is processed assuming that drugs of this category *R05D2* may all have the *productive cough* contra-indication. This aspect corrects data quality completeness related to comission. It is also possible to correct data quality completeness related to omission but deleting contra-indications in the SPC drug window.

- a product-centric approach where for a given drug or group of drugs, defined on a non-related to the ontologies criteria (e.g. therapeutic classes, phramaceuticals), a matrix for each SPC field specified is displayed, e.g. one for contra-indication, another one for side-effects, etc. These matrices are similar in principle to the one described above but it proposes interactions with both ontologies at the same time. Figure 4 proposes an extract for the contra-indication SPC field for the *antitussive* therapeutic class. In such a matrix, we are using colors and integer values to highlight violation candidate cells. The integer values in cells of a matrix correspond to:

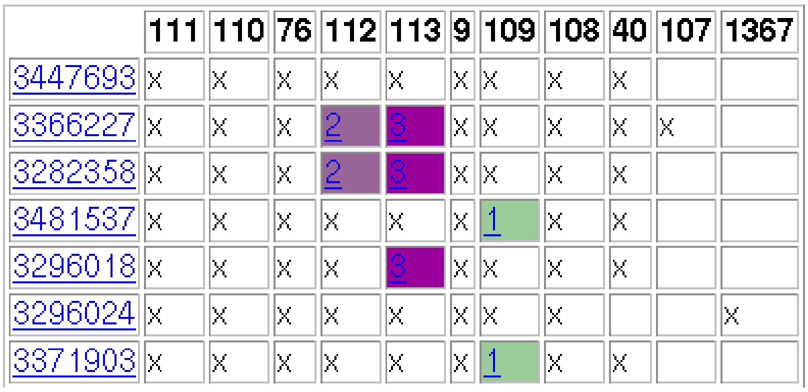
  - A cross character ('x') when the drug already contains this contra-indication.
  - A value of 1 in a cell highlights a proposition made from the EphMRA ontology. This is the case for the contra-indication with value 109 and products identified with CIP 3481537 and 3371903.
  - A value of 2 in a cell highlights a detection made from inferences using the ATC ontology.
  - A value of 3 in a cell highlights that both ontologies (EphMRA and ATC) have detected this cell as a candidate for violation.
  - A blank character when the drug does not present the contra-indication and nothing has been detect for this product from either of both ontologies.

The main assumption behind this method is that a column of crosses (column 111 on Figure 4) highlights that all drugs of this group present this contra-indication. On the other-hand, a column with very few crosses (column 107 on Figure 4) may indicate a completeness comission which can be easily repaired from the drug page by simply removing the contra-indication. Problems related with completeness omission are dealt with colored/valued cells. If a cell displays a value of 3, it means that reasoning on both ontologies indicates a violation, and thus the violation detection may be relevant.

4.2 Evaluation

In this section ,we are interested in evaluating the effectiveness of our inductive approach and also to evaluate to user-friendliness and efficiency of the detection and repairing graphical user interface. We propose to evaluate these aspects in terms of: (1) improvement of the resulting drug database after a thorough detection and repairing step, (2) satisfaction of the team of health care professionals maintaining the database. The evaluation emphasizes the results of the first execution of the detection and repairing process. This is the most relevant results as it is the step where the enhancements are most clearly visible. As database updates are performed, the data quality improvements are less evident but still as effective.

The first evaluation aspect of this method is the precision of the database obtained after a thorough repairing of the drug database. On the SPC fields studied (contra-indications, drug interactions, cautions with allergies and diseases, cautions with other drug treatments, side-effects), we improved the precision of the drug database by 8%. This precision takes into account, the total number of tokens (terms in any of previously listed SPC fields) added, deleted, the total number of tokens before and after repairing. This



	111	110	76	112	113	9	109	108	40	107	1367
3447693	x	x	x	x	x	x	x	x	x		
3366227	x	x	x	2	3	x	x	x	x	x	
3282358	x	x	x	2	3	x	x	x	x		
3481537	x	x	x	x	x	x	1	x	x		
3296018	x	x	x	x	3	x	x	x	x		
3296024	x	x	x	x	x	x	x	x			x
3371903	x	x	x	x	x	x	1	x	x		

Figure 4: Extract from a product-centric view : contra-indication matrix for antitussive therapeutic class

rate of improvement shows that the original database was of an unsurprisingly high quality. In fact, we were confident about the relative quality of our database and this was the main motivation to use this inductive reasoning approach. Anyhow, this result also shows that this database's data quality can be improved by our method.

The second aspect which is quite interesting is the satisfaction of our health care professionals. They consider that the detection and repairing method is really user-friendly, as it only involves reading and clicking, and the learning is quite short, as the detection only requires a visual detection. The best proof of success is the high usage rate of this maintenance assistant by our team.

The efficiency of the detection method depends on the quality of the grouping factor. We can stress that the fewer products in the group, the easier it is to reach the membership threshold θ without really being relevant.

5. COMPOUND DRUG MODELIZATION ENHANCEMENT

We have seen in the previous section that the evaluation of the method is less efficient for groups that contain few products or for groups that are associated with ontology concepts, meaning EphMRA or ATC codes, that are not really relevant in the medical domain. A typical example of the latter is the presence of codes corresponding to compound drugs. In Sections 2.2.1 and 2.2.2, we faced such a codes in both EphMRA and ATC terminologies, i.e. *R5D2* and *R05DA20*. In this rest of this section, we study a solution for the ATC system which is also relevant for EphMRA.

Using a compound concept in the ATC-ontology results in grouping drugs that may not have much in common. This issue is important as 475 codes out of the 5309 codes of the ATC classification are relevant to compound codes. A concrete example highlights the problems encountered. The products such as *Camphodionyl©* , made of *codein* and *sulfoguaiacol*, and *Hexapneumine©* , made of *pholcodin* and *chlorphenamin* and *biclotymol*, do not have a single active molecule in common. Thus inferences using the ATC ontology will likely not be effective. In order to avoid such problems, we propose to classify these drugs with the conjunction of their molecule. For example, the code of *Camphodionyl©* will be *R05DA04* (for *Codein*) and R05CA09 (for *sulfoguaiacol*). Using this classification method it is now possible to detect violations for compound drugs using in this example, the associated terms from both *Codein* and *sulfoguaiacol*.

6. CONCLUSION

In this paper we presented a simple yet effective solution to enhance the data quality of drug databases. The method exploits inductive reasoning to provide terminologies with some formal presentation of definitional drug information. These ontologies are later exploited to semi-automatically detect and automatically repair ontology-based integrity constraints violations. A particular attention was given on the user-friendliness aspect of the detection and repairing graphical interface. An evaluation on a self-medication database showed that the method enables to improve the content of database instances. We believe that this inductive approach can be generalized to other domains where terminologies/ontologies and quality databases are accessible.

The detect and repair approach can be improved on several aspects. For instance, we are working on a solution that automatically sets a maximized threshold θ for couple of SPC fields and ontologies. In the current solution, the thresholds are defined manually and empirically by health care professsionals. Such a task can be considered laborious and imprecise.

Another extension consists in automatizing the generation of SPCs from our newly designed and enriched ontologies. This task needs the integration of additional drug terminologies such as the Defined Daily Dose (DDD), generally associated to the ATC, in order to define drug posologies.

Finally, the enrichment of terminologies with terms coming from a given database enables to support ontology mappings. We are currently studying the potential and benefits of such an approach in the context of drug databases and terminologies.

7. REFERENCES

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison Wesley, 1995.

[2] M. Ball, J. Lillis. E-health : transforming the physician-patient relationship. *International Journal of Medical Informatics*, 61:1-10, 2001.

[3] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P.F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.

[4] A. Borgida. On the relative expressiveness of Description Logics and precidate logics. *Artificial intelligence*, 82(1-2):353-367, 1996.

[5] J.J. Cimino, X. Zhu. The practical impact of ontologies on biomedical informatics *IMIA Yearbook of Medical Informatics*:1-12, 2006.

[6] O. Curé, R. Squelbut. A database trigger strategy to maintain knowledge bases developed via dat a migration. *Proceedings of EPIA 2005. LNAI 3808*:206-217, 2005.

[7] O. Curé. Ontology Interaction with a Patient Electronic Health Record *Proceedings of 18th IEEE Symposium on Computer-Based Medical Systems*:185-190, 2005.

[8] O. Curé, R. Squelbut. Integrating data into an OWL Knowledge Base via the DBOM Protégé plug-in. *Proceedings of the 9th International Protégé conference*, 2006.

[9] M. Dean, G. Schreiber. OWL Web Ontology Language Reference. *W3C Recommendation*, 2004.

[10] J. de Bruijn, R. Lara, A. Polleres, and D. Fensel. OWL DL vs. OWL flight: conceptual modeling and reasoning for the semantic Web. *Proceedings of 14th international conference on World Wide Web*: 623–632, 2005.

[11] J. Gennari, M. Musen, R. Fergerson, W. Grosso, M. Crubezy, H. Eriksson, N. Noy, S. Tu. The evolution of protege: an environment for knowledge - based systems development. *International Journal of Human - Computer Studies* 123:58-89, 2003.

[12] J.P. Giroud, C. Hagège. *Le guide Giroud-Hagège de tous les médicaments* Editions du Rocher, Paris France. 2001.

[13] P.C. Kanellakis. Elements of relational database theory *Handbook of theoretical computer science (vol. B): formal models and semantics*:1073–1156, MIT Press, 1990.

[14] Bridging the gap between OWL and relational databases. *Proceedings of the 16th International World Wide Web Conference*, to appear, 2007.

[15] R. Reiter. What Should a Database Know? *Journal of Logic Programming*, 14(1-2):127-153, 1992.

[16] E.Sirin, B. Parsia, B. Cuenca Grau, A. Kalyanpur, and Y. Katz. Pellet: A Practical OWL-DL reasoner. *Journal of Web Semantics* To appear.

[17] WHO Collaborating Centre for Drug Statistics Methodology URL of Web site : http://www.whocc.no/atcddd/