MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190187

Romedi: An Open Data Source About French Drugs on the Semantic Web

Sébastien Cossin[ab], Luc Lebrun[b], Grégory Lobre[a], Romain Loustau[a], Vianney Jouhet[ab], Romain Griffier[a], Fleur Mougin[b], Gayo Diallo[b], Frantz Thiessard[ab]

[a]Bordeaux university hospital, Pôle de santé publique, Service d'information médicale, Unité Informatique et Archivistique Médicales, F-33000 Bordeaux, France
[b]Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, team ERIAS, UMR 1219, F-33000 Bordeaux, France

Abstract

The W3C project, "Linking Open Drug Data" (LODD), linked several publicly available sources of drug data together. So far, French data, like marketed drugs and their summary of product characteristics, were not integrated and remained difficult to query. In this paper, we present Romedi (Référentiel Ouvert du Médicament), an open dataset that links French data on drugs to international resources. The principles and standard recommendations created by the W3C for sharing information were adopted. Romedi was connected to the Unified Medical Language System and DrugBank, two central resources of the LODD project. A SPARQL endpoint is available to query Romedi and services are provided to annotate textual content with Romedi terms. This paper describes its content, its services, its links to external resources, and expected future developments.

Keywords:

Pharmaceutical Preparations
Semantics
Vocabulary, Controlled

Introduction

Drug information is spread over multiples sites on the Internet. Summary of Product Characteristics (SPCs) are documents produced by pharmaceutical companies and approved by public health agencies. In France, ANSM (Agence Nationale de Sécurité du Médicament et des Produits de Santé), the French Medicines Agency, publishes SPCs approved by itself or the European Medicines Agency (EMA) on a website. An SPC contains key information about a marketed drug like the therapeutic indication(s), the posology, dosage adjustment, drug-drug interactions, and contraindications [1,2].

However, other additional knowledge related to marketed drugs can be found on the Internet and is not clearly connected. For example, the French thesaurus of drug-drug interaction (DDI), edited by ANSM, is the official reference document on this topic. This document, available as a PDF file, describes potential DDI (PDDI) between molecules. The links between molecules in the SPC and molecules in the reference document are not explicit and cannot be linked automatically due to semantic and syntactic interoperability issues. Another example is information about drugs' safety during pregnancy. The CRAT [3] (Centre de référence sur les agents tératogènes) is a French public organization especially involved in this public health issue. It provides free access to information about risks of drug intake during pregnancy that often disagrees with the SPC documents [2]. Still, connections and comparisons between these two sources can only be made by humans.

Furthermore, specific or general international sources, like DrugBank [4] and DBpedia [5], deliver supplementary information about drugs marketed abroad or about characteristics of molecules. DrugBank is a comprehensive, freely accessible, online database containing a large amount of information on molecules (e.g., chemical structure, half-life). DBpedia provides structured, machine-understandable knowledge extracted from Wikipedia articles. Many drugs and molecules are described by DBpedia contributors. The Unified Medical Language System [6] (UMLS) is a compendium of a large number of national and international vocabularies. In particular, UMLS contains RxNorm [7], a standard nomenclature developed by the United States National Library of Medicine (NLM) in the field of medications and the MeSH [8], a comprehensive controlled vocabulary for the purpose of indexing scientific articles.

The scattering of information is a significant issue for information retrieval which hampers the reusability of up-to-date knowledge on the web. Tim Berners-Lee, the inventor of the World Wide Web, suggested a 5-star deployment scheme for sharing information on the web (figure 1).

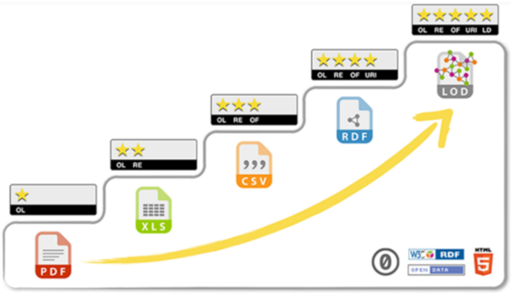

Figure 1– 5-star deployment scheme for Open Data [1].

The idea is to use W3C standards when publishing data in order to create a semantic graph that is capable of interlinking information of various datasets distributed over the web.

The W3C project, "Linking Open Drug Data" (LODD), focuses specifically on linking various sources of drug data together [9]. Participants of the LODD project have already made dozens of datasets relevant to pharmaceutical research and development available as linked data. In this paper we present Romedi, a new open dataset on French drugs linked to other related resources on the semantic web. In the methods section we discuss the Romedi data model and how it was linked to external resources. In the results section, examples of use cases are described.

[1] Source : https://5stardata.info/en/

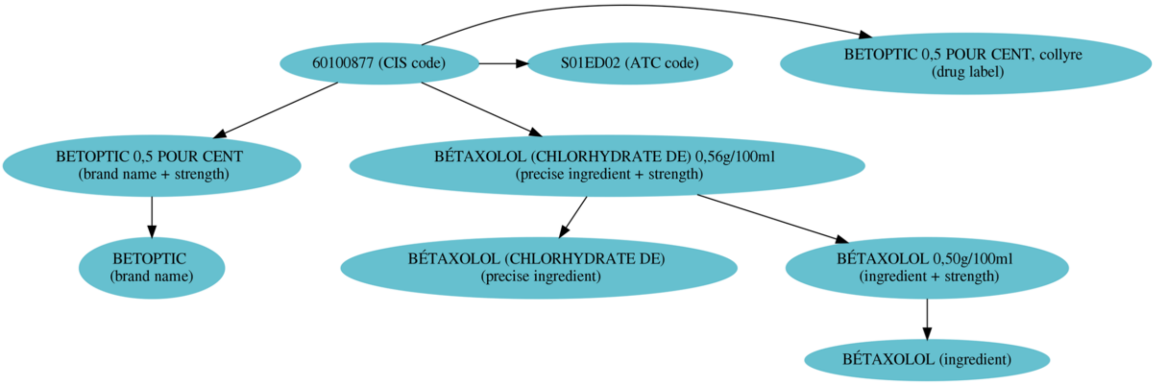

Figure 2– The Romedi data model instantiated with the drug "BETOPTIC 0.5 POUR CENT, collyre". The normalization step extracted the brand name (BETOPTIC) and the strength from the drug label. The ingredient was also normalized and linked to its corresponding UMLS concept C0005320 and DrugBank concept DB00195

Methods

Data Model

The Romedi data model is close to the RxNorm terminology [7]. It contains similar classes like the brand name (BN), ingredient (IN), and precise ingredient (PIN). The CIS code is an identifier of a marketed drug in France, and the URL (Uniform Resource Locator) to access the SPC depends on this code. For example, "BETOPTIC 0.5 POUR CENT, collyre" (Figure 2), is the label of a drug identified by the CIS code 60100877. The SPC of this drug is accessible at http://base-donnees-publique.medicaments.gouv.fr/affichageDoc.php?specid=60100877&typedoc=R. Each marketed drug has one or several ATC codes. The Anatomical Therapeutic Chemical (ATC) is a widely used system of alphanumeric codes developed by the World Health Organization (WHO)[2] for the classification of drugs.

In France, the main data source for marketed drugs is the "base de données publique des médicaments", a freely accessible database available in a text file format[3] which is updated every month by national health authorities. It contains all currently marketed drugs and drugs withdrawn from market in the last three years. The database contains details on marketed drugs like CIS codes, drug labels, and molecules. A normalization and transformation process is needed to instantiate the Romedi model as the concept of brand name is not present in this database. For example, "INEXIUM 20mg, comprimé gastro-résistant" is a drug label but the brand name "INEXIUM" is not present and must be extracted. In addition, only the precise ingredient is present for some drugs. For example, "pravastatine sodique" appears but the term "pravastatine", the ingredient, is missing. This normalization step was done by using regular expressions algorithms and an interface for manual validation by a pharmacist. This step was fully described elsewhere [10].

After model instantiation, ingredient instances were linked to UMLS [6] (Unified Medical Language System) and DrugBank [11]. The mapping between the French and international reseources was done as follows: the mapping is automatic if two terms have a perfect match, semi-automatic with a validation interface when a partial match is found, and manual when a French term cannot be found.

Ingredients were also linked to the French thesaurus of DDI. The automatic extraction and transformation of the PDF document to a CSV file was done with an R package[4]. The first use case to exploit external resources links was to compare French reference document on DDI with international ones. Ayvaz et al. [12] managed to gather information about DDI from publicly available sources, and the authors used DrugBank identifiers to describe couples of DDI interaction from different sources. The authors made a merged-PDDI database that can be downloaded online. The external links to DrugBank were used to integrate French knowledge to the merge PDDI database and to automatically compare the French national reference with other sources.

Medication Extraction Module

Automated identification of drugs in unstructured data, like social media or clinical notes, is essential for post-marketing drug safety surveillance that aims to detect signals of drug misuse or adverse event effects [13,14]. The natural language processing goal is not only to identify the terms in a corpus that correspond to drug entities, but also to map these entities to a well-established knowledge base. 'Semantic annotation' is the name given to this task by Jovanovic et al. [15].

IAMsystem, a general semantic annotation tool, was developed to facilitate the identification of Romedi terms in textual content. It was initially developed for the DoMINO project (Drug Misuses In Networks) that aims to detect drug misuse in fora [16]. The program performances were evaluated on a shared task for disease detection using death certificates [17], and the program was described in-detail at this occasion [18]. IAMsystem is open-source and available on GitHub[5]. It takes a set of terms as inputs, normalizes them, and stores them in a tree data structure for fast dictionary look-up. It handles abbreviations, a set of which were manually added for drug detection (e.g., "ac" for "acide", "vit" for "vitamine"). The typo module for drug detection is a logistic regression trained on a manually created gold standard of 3,438 potential spelling errors of brand names and molecules. Three explanatory variables are used:

[2] https://www.whocc.no/atc_ddd_index

[3] http://base-donnees-publique.medicaments.gouv.fr

[4] https://github.com/scossin/IMthesaurusANSM

[5] https://github.com/scossin/IAMsystem

1. The length of the potentially misspelled word
2. The similitude of the first letter between the potential misspelled word and the dictionary word
3. Levenshtein's distance between the phonetic transformation of two words

This last variable is computed with the French "phonetic" algorithm of the Talisman program[6]. The model is able to predict a typo of a brand name or an ingredient with a specificity of 0.93 and a sensitivity of 0.60. The performances of the annotator were evaluated in clinical texts, and the results are presented in the following section.

Romedi terms can be used to detect French drugs in textual content. Links to the CRAT website were established by detecting brand names in the alphabetic index of the web page. The graph model permits easy retrieval of information about the detected drug , such as the ATC code(s) or external resources links.

Results

Romedi Content

The first version of the Romedi terminology contains 13,661 French marketed drugs, 4,277 brand names, and 2,109 ingredients after the normalization step. Among the ingredients, 1,918 (91%) were linked to a UMLS concept and 1,434 (71%) to a DrugBank concept. 954 (95%) of the molecules in the French DDI thesaurus were mapped to a Romedi ingredient or precise ingredient. Since DrugBank contains French synonyms of ingredients and the French version of MeSH is integrated in the UMLS, most of the mappings were done automatically.

A web interface that makes the Romedi content available is accessible at https://www.romedi.fr. Like RxNav [19], the interface displays links between clinical drugs, active ingredients, brand names, ATC code(s), and external resources. Drug information from external resources is retrieved using SPARQL queries or application programming interfaces (API). For example, the DBpedia definition of a molecule is retrieved by a SPARQL query to its endpoint[7] and by using DrugBank links to DBpedia. Romedi resources are identified by their Uniform Resource Identifiers (URIs). For example, https://www.romedi.fr/romedi/CIS60100877 is the URI of the Romedi drug "BETOPTIC 0.5 POUR CENT, collyre", and is also a valid URL. Retrieving a representation of a resource identified by a URI is known as dereferencing a URI, and it can be used to obtain a representation that can be perceived by a user.

A SPARQL endpoint[8] is also available to query Romedi content. In addition, the terminology is freely downloadable as an RDF file and can be reused under an open license.

Automatic Comparison of the French DDI Reference Document with International

Mapping national ingredient concepts to international ones allows users to automatically compare drug-drug interaction information. Applied on more than 7 million drugs deliveries in France, the main discrepancy between the French thesaurus and international sources was with the couple "escitalopram – flecainide." This drug pair was considered contraindicated by an international source, and no risk of interaction is described in the French thesaurus although it contains four levels of severity. Full results of this work are available elsewhere [20]. The French thesaurus was integrated in the merged-PDDI dataset[9] in a linked open format.

Medication Extraction Module

Brand names and molecule identification performances were evaluated using the 'current medication' section in electronic health records (EHR) form from Bordeaux Hospital's emergency department. Among the 6,070 drugs detected (brand names or molecules), the specificity/sensitivity were 0.99/0.92 and 0.99/0.96 without and with the typo module respectively [10]. The performance of these models can be explained by the grocery list type of the input data and disregarding medication attributes (e.g., dosage, strength, and route). Further evaluation is required in narrative clinical notes. The programs can be installed locally[10] and the annotator is also available in an R package[11].

Discussion

Open Data has the potential to provide significant benefits to society. The Link Data movement aims to share and integrate knowledge on the Web. Still, many resources containing important information on drugs remain inaccessible to machine and hampers automatic comparison of knowledge.

Romedi is a French open dataset that connects French and international information resources about drugs.

First, the French dataset on marketed drugs (base de données publique des médicaments) was normalized and integrated in an RxNorm-like data model. Then, national and international sources were linked, including the French thesaurus of PDDI, the French reference for evaluating drugs safety during pregnancy (CRAT), DrugBank, and UMLS. An interface was developed to navigate Romedi terminology, and a SPARQL endpoint is accessible to query its content.

Developers can use Romedi API to retrieve information about French drugs for their website. Researchers can use Romedi services to extract drugs in textual content for post-marketing drug safety surveillance. Using the same terminology, data can be shared with other researchers to promote collaboration and enhance pharmacovigilance discovery.

Romedi URIs are currently used to index drugs in clinical notes in Bordeaux Hospital's i2b2 [21] data warehouse and to ease information retrieval. Searches can be performed by brand name, molecule, or ATC code. Medication extraction and indexation helps patient phenotyping and cohort identification tasks.

Future Developments

The current version of Romedi is 2.2 and the updating process is not fully automated. The biggest challenge will be maintainance of this terminology, especially correcting errors identified by users and managing updates of French-marketed drugs and external resources. Our aim is to build an engaged community of users and developers around Romedi.

So far, the Romedi data model contains no description logics. It would be feasible to infer logical equivalences between Romedi and RxNorm drugs by using a formal model. A mapping between Romedi and RxNorm beyond the ingredient level would facilitate French drug data integration in the Observational Health Data Sciences and Informatics (OHDSI) common data model that uses RxNorm as an international standard [22]. This mapping will be important to involve French researchers in international collaborations and research

[6] https://github.com/Yomguithereal/talisman/tree/master/src/phonetics

[7] https://dbpedia.org/sparql

[8] http://www.romedi.fr:8890/sparql

[9] https://www.dikb.org/Merged-PDDI/

[10] https://github.com/scossin/RomediApp

[11] https://github.com/scossin/RIAMsystem/

projects on drugs, and it may help our national agency to exchange information about post-marketing surveillance.

Conclusions

In this paper we have presented Romedi, an open dataset about French drugs linked to multiple resources on the semantic web. Web services are provided to annotate textual documents and query Romedi content to retrieve additional information about detected drugs.

Acknowledgements

This project is part of the Drugs Systematized Assessment in real-liFe Environment (DRUGS-SAFE) research platform, funded by the French Medicines Agency (Agence Nationale de Sécurité du Médicament et des Produits de Santé, ANSM). This platform aims to provide an integrated system allowing the concomitant monitoring of drug use and safety in France.

References

[1] M. Rougemont, S. Ulrich, C. Hiemke, E. Corruble, and P. Baumann, French summaries of product characteristics: Content in relation to therapeutic monitoring of psychotropic drugs, *Fundamental & Clinical Pharmacology* **24** (2010), 377–384.

[2] B. Arguello, T.M. Salgado, and F. Fernandez-Llimos, Assessing the information in the Summaries of Product Characteristics for the use of medicines in pregnancy and lactation, *Br J Clin Pharmacol* **79** (2015), 537–544.

[3] E. Elefant, C. Vauzelle, and D. Beghin, [Centre de référence sur les agents tératogènes (CRAT): A pioneer center], *Therapie* **69** (2014), 39–45.

[4] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A.C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z.T. Dame, B. Han, Y. Zhou, and D.S. Wishart, DrugBank 4.0: Shedding new light on drug metabolism, *Nucleic Acids Res* **42** (2014), 1091–1097.

[5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, DBpedia: A Nucleus for a Web of Open Data, in: Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference, Springer-Verlag, Berlin, Heidelberg, 2007: pp. 722–735.

[6] O. Bodenreider, The Unified Medical Language System (UMLS): Integrating biomedical terminology, *Nucleic Acids Res* **32** (2004), 267-270.

[7] S.J. Nelson, K. Zeng, J. Kilbourne, T. Powell, and R. Moore, Normalized names for clinical drugs: RxNorm at 6 years, *J Am Med Inform Assoc* **18** (2011), 441–448.

[8] S. Kim, L. Yeganova, and W.J. Wilbur, Meshable: Searching PubMed abstracts by utilizing MeSH and MeSH-derived topical terms, *Bioinformatics* **32** (2016), 3044–3046.

[9] M. Samwald, A. Jentzsch, C. Bouton, C.S. Kallesøe, E. Willighagen, J. Hajagos, M.S. Marshall, E. Prud'hommeaux, O. Hassenzadeh, E. Pichler, and S. Stephens, Linked open drug data for pharmaceutical research and development, *J Cheminform* **3** (2011), 19.

[10] S. Cossin, R. Loustau, V. Jouhet, F. Mougin, G. Evrard, C. Giljardine, G. Diallo, and F. Thiessard, ROMEDI, une terminologie médicale française pour la détection des médicaments en texte libre, Artificial Intelligence Platform, Nancy (2018).

[11] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, DrugBank: A comprehensive resource for in silico drug discovery and exploration, *Nucleic Acids Res* **34** (2006), 668-672.

[12] S. Ayvaz, J. Horn, O. Hassanzadeh, Q. Zhu, J. Stan, N.P. Tatonetti, S. Vilar, M. Brochhausen, M. Samwald, M. Rastegar-Mojarad, M. Dumontier, and R.D. Boyce, Toward a complete dataset of drug-drug interaction information from publicly available sources, *J Biomed Inform* **55** (2015), 206–217.

[13] C.L. Ventola, Big Data and Pharmacovigilance: Data Mining for Adverse Drug Events and Interactions, *P T.* **43** (2018), 340–351.

[14] S. Sohn, C. Clark, S.R. Halgrim, S.P. Murphy, C.G. Chute, and H. Liu, MedXN: An open source medication extraction and normalization tool for clinical text, *J Am Med Inform Assoc* **21** (2014), 858–865.

[15] J. Jovanović, and E. Bagheri, Semantic annotation in biomedicine: the current landscape, *J Biomed Semantics* **8** (2017).

[16] E. Bigeard, Construction de lexiques pour l'extraction des mentions de maladies dans les forums de santé, *19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL)* (2017), 15-27

[17] A. Névéol, A. Robert, F. Grippo, C. Morgand, C. Orsi, L. Pelikan, L. Ramadier, G. Rey, and P. Zweigenbaum, CLEF eHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian, in: CLEF, 2018.

[18] S. Cossin, V. Jouhet, F. Mougin, G. Diallo, and F. Thiessard, IAM at CLEF eHealth 2018: Concept Annotation and Coding in French Death Certificates, *ArXiv:1807.03674 [Cs]* (2018).

[19] K. Zeng, O. Bodenreider, J. Kilbourne, and S.J. Nelson, RxNav: A Web Service for Standard Drug Information, *AMIA Annu Symp Proc* (2006), 1156.

[20] S. Cossin, Intéractions médicamenteuses, données liées et applications, Thèse d'exercice, Université de Bordeaux, 2016.

[21] S.N. Murphy, M.E. Mendis, D.A. Berkowitz, I. Kohane, and H.C. Chueh, Integration of clinical and genetic data in the i2b2 architecture, *AMIA Annu Symp Proc* (2006), 1040.

[22] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Y.-C. Li, P.E. Stang, D. Madigan, and P.B. Ryan, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers, *Stud Health Technol Inform* **216** (2015), 574–578.

Address for correspondence

sebastien.cossin@u-bordeaux.fr