Ontologies biomédicales et Web Sémantique pour la réutilisation des bases de données médico-administratives en pharmaco-épidémiologie

Yann Rivault, Olivier Dameron, Nolwenn Le Meur

Ontologies biomédicales et Web Sémantique pour la réutilisation des bases de données médico-administratives en pharmaco-épidémiologie

Y. Rivault[1,2] — O. Dameron[2] — N. Le Meur[1]

[1] Univ Rennes, EHESP, REPERES (Recherche en Pharmaco-Épidémiologie et Recours aux Soins) - EA 7449, F-35000 Rennes, France
yann.rivault@ehesp.fr - nolwenn.lemeur@ehesp.fr
[2] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France
olivier.dameron@univ-rennes1.fr

RÉSUMÉ. *Les bases de données médico-administratives françaises sont de plus en plus réutilisées pour la recherche en santé publique, par exemple en pharmaco-épidémiologie et pharmacovigilance. L'exploration de ces données volumineuses et hétérogènes, dans le cadre de ces nouveaux usages, est loin d'être triviale. Une approche basée sur l'utilisation d'ontologies et des technologies du Web Sémantique permet de lier des données patient à des connaissances médicales et pharmacologiques, et offre ainsi de nouvelles opportunités d'exploration pour les données médico-administratives. Néanmoins, cette approche n'est que rarement utilisée avec ces données. Nous présentons cette approche ainsi qu'un package R, queryMed, qui vise à favoriser l'utilisation d'ontologies médicales et pharmacologiques par les épidémiologistes et biostatisticiens.*

ABSTRACT. *French health care administrative databases are increasingly reused for public health research, for example in pharmacoepidemiology and pharmacovigilance. Exploring these massive and heterogeneous data as part of these new usages is far from trivial. An approach based on the use of ontologies and Semantic Web technologies allows linking patient data with medical and pharmacological knowledge. It thus offers new opportunities to explore patient data issued from health care administrative databases. Nevertheless, this approach is rarely used in this field. We present this approach as well as an R package, queryMed, which aims to promote the usage of medical and pharmaoclogical ontologies by epidemiologists and biostatisticians.*

MOTS-CLÉS : *Bases de données médico-administratives, Ontologies, Web Sémantique, Pharmaco-épidémiologie, SNDS*

KEYWORDS: *Health care administrative databases, Ontologies, Semantic Web, pharmacoepidemiology, SNDS*

1. Introduction

En France, la réutilisation des bases de données médico-administratives, regroupées au sein du Système National des Données de Santé (SNDS), ouvre de nouvelles perspectives pour la recherche en santé publique. En pharmaco-épidémiologie, le traitement de telles données permet ainsi d'étudier au niveau populationel, l'état de santé, les maladies ainsi que la consommation et le recours aux soins[ROU 18]. En pharmaco-vigilance, ces données permettent de surveiller les effets secondaires, les interactions et contre-indications associées aux médicaments. Ces nouvelles utilisations et les complexités de ces systèmes d'information, soulèvent des défis en terme de représentation, d'intégration et d'exploration des données. Dans le domaine des sciences de la vie, de nombreux travaux ont montré que de tels défis pouvaient être relevés grâce à l'utilisation des technologies du Web Sémantique et d'ontologies du domaine[STE 06]. Par ailleurs, des outils ont contribué à rendre les ontologies et technologies du Web Sémantique plus accessibles aux statisticiens depuis le logiciel de statistique et langage de programmation R [Van 13, WIL 14, KUR 15]. Cependant, ces outils, technologies et approches, ne sont que peu utilisés dans le domaine de la santé publique, plus spécifiquement en pharmaco-épidémiologie [FER 13]. A notre connaissance, aucun de ces outils n'est vraiment spécifique aux connaissances médicales et pharmacologiques, à leurs réutilisations pour l'exploration, puis l'analyse statistique, de données médico-administratives. Nous présentons en premier lieu dans cet article une approche de représentation, d'intégration et d'exploration des données médico-administratives fondée sur l'utilisation des technologies du Web Sémantique et l'apport d'ontologies biomédicales. Puis nous présentons queryMed[1], un package R [R C 17] qui vise à rendre cette approche plus accessible pour les chercheurs en santé publique en facilitant l'intégration de connaissances issues des principales ontologies médicales et pharmacologiques.

2. Représentation, intégration et exploration des bases de données médico-administratives fondées sur l'utilisation des technologie du Web Sémantique et l'apport d'ontologies biomédicales

La multitude des sources de recueil de données mène à une hétérogénéité et à une volumétrie importante des bases médico-administratives. Parallèlement, leur exploration est loin d'être triviale. En plus d'une architecture très complexe, leur exploration nécessite la connaissance des nomenclatures médicales et comptables qui les régissent. Les diagnostics réalisés à l'hôpital sont par exemple codifiés selon la Classification Internationale des Maladies - $10^{\text{ème}}$ révision (CIM-10), une codification hiérarchique constituée de plus de 14 000 codes. Si cette codification systématique est souvent vue comme une complexité du fait de la multitude et de la profondeur des nomenclatures, elle représente un réel atout en supportant l'interopérabilité sémantique et ainsi le lien entre données et ontologies médicales. Dans le cadre de la réutilisation de ces données

1. `https://github.com/yannrivault/queryMed`

pour la recherche en pharmaco-épidémiologie, nous proposons d'utiliser les technologies du Web Sémantique et des ontologies biomédicales pour représenter, intégrer et explorer ces données.

Dans cette approche, les données patient peuvent être représentées selon un modèle de graphe, le Ressource Description Framework[2] (RDF). Des nomenclatures présentes dans ces systèmes d'information, telles que la CIM-10, la classification Anatomique Thérapeutique et Chimique (ATC) et la Classification Commune des Actes Médicaux (CCAM), décrites aux format du Web Sémantique sont reliées au graphe RDF des données patient. Des correspondances entre les termes de ces nomenclatures et ceux d'ontologies permettent alors de lier des données patient à tout un ensemble de connaissances médicales et pharmacologiques plus riches, issues d'ontologies du Linked Data. Les Concept Unique Identifier (CUI) de l'Unified Medical Language System[3] (UMLS) permettent une correspondance entre les codes de diagnostic CIM-10 et les codes de médicament de l'ATC à ceux de la National Drug File - Reference Terminology[4] (NDF-RT), une ontologie décrivant médicaments et diagnostics, certains effets physiologiques de ces médicaments, leurs contre-indications, ou encore leurs mécanismes d'action. Les codes ATC peuvent aussi être reliés aux codes des médicaments de DrugBank[5], une base de données décrivant les médicaments, leurs potentielles interactions et leurs caractéristiques chimiques. Parallèlement, ces correspondances permettent de réutiliser des travaux visant à regrouper certaines ressources du Linked Data décrivant les médicaments. La Drug Indication Database (DID) [SHA 17], une base de données regroupant des connaissances issues de douze sources, dont certaines du Linked Data, décrit ainsi les potentielles indications de médicaments. De la même façon, la Drug Interaction Knowledge Base (DIKB) [AYV 15] regroupe quatorze sources de connaissances pour décrire les interactions potentielles des médicaments. Codifiés en CUI et en code DrugBank, les médicaments, les diagnosctics, les indications et les intérractions qui y sont décrits peuvent alors être reliés aux nomenclatures de l'ATC et de la CIM-10. Tous ces liens entre données et ontologies ou autres systèmes d'organisation de la connaissance du domaine médical et pharmacologique permettent alors d'explorer les données à travers les nœuds et arcs d'un graphe.

Dans cette approche, l'exploration de ce graphe est réalisée avec SPARQL[6] (SPARQL Protocol And Querry Language), la technologie du Web Sémantique pour le requêtage des données et connaissances représentées dans ses standards. Cette approche permet de répondre à des questions relativement complexes en pharmaco-épidémiologie, auxquelles il est difficile de répondre sans l'ajout de connaissances du domaine :

– Quels patients présentent un lien d'indication entre médicament et diagnostic ?

2. https://www.w3.org/RDF/
3. https://www.nlm.nih.gov/research/umls/
4. https://bioportal.bioontology.org/ontologies/NDFRT
5. https://www.drugbank.ca/
6. https://www.w3.org/TR/rdf-sparql-query/

– Quels patients présentent une contre-indication médicament-diagnostic ?
– Quels patients présentent une interaction médicamenteuse ?
– Quels patients présentent une contre-indication médicamenteuse ?

3. Favoriser la réutilisation des connaissances médicales pour la pharmaco-épidémiologie : queryMed

Si de plus en plus de connaissances médicales sont accessibles via le Web des données et peuvent être reliées entre elles et aux données de santé grâce à des efforts de correspondances, leur réutilisation reste une tâche compliquée pour les non-initiés aux technologies du Web Sémantique ou aux ontologies. Pour pallier à cette difficulté, nous développons le package R queryMed. Son objectif est d'aider les chercheurs dans l'utilisation de connaissances médicales et pharmacologiques issues du Linked Data, ainsi que dans la liaison de ces connaissances aux données de santé, plus particulièrement aux données provenant des bases médico-administratives. queryMed propose des fonctions de requêtage pour les SPARQL endpoints, des serveurs mettant à disposition des données et connaissances aux standards du Web Sémantique, requêtable grâce à SPARQL. Le package offre des requêtes prédéfinies adaptées aux SPARQL endpoints et au domaine médical ainsi que pharmacologique : Bioportal [WHE 11], SIFR BioPortal [JON 16], Bio2rdf [CAL 13], DB-pedia [LEH 15] et Ontobee [ONG 17]. Ces requêtes permettent de récupérer des connaissances à propos de codes de médicaments et de diagnostics, par exemple des définitions, des labels, des indications de médicaments, des contre-indications, des interactions ou encore des effets physiologiques. queryMed intègre également les bases de connaissances DID et DIKB. Des fonctions permettant d'utiliser des correspondances rend possible la liaison de toutes ces ressources à des données de santé codifiées, par exemple des codes de diagnostics CIM-10 ou des codes de médicaments de l'ATC. Enfin, le package offre des fonctions permettant de rechercher des relations entre codes de médicaments et diagnostics tels les indications, les contre-indications ou encore les interactions médicamenteuses, pour des ensembles de patients.

4. Application

Nous avons utilisé queryMed pour explorer les prescriptions de médicaments de 1003 patients opérés d'une angioplastie pour artériopathie oblitérante des membres inférieurs (AOMI), en France, en 2015. Les prescriptions de médicaments 15 jours avant et après la chirurgie ont été extraites du SNDS pour chacun de ces patients. Nous nous sommes intéressés à identifier, sur la base de connaissances issues d'ontologies, les patients dont les consommations de soins comportaient une contre-indication entre médicament et diagnostic, ou un lien d'indication entre médicament et diagnostic.

Le package a permis de relier les codes ATC et CIM-10 de ces patients aux connaissances médicales et pharmacologiques de NDF-RT et DID. L'ajout de ces connais-

sances nous ont permis d'identifier un patient présentant une contre-indication médicamenteuse avec son AOMI, ainsi que 931 patients avec au moins un prescription de médicament indiqué pour l'AOMI, soit 72 sans aucune prescription relative à l'AOMI (Figure 1).

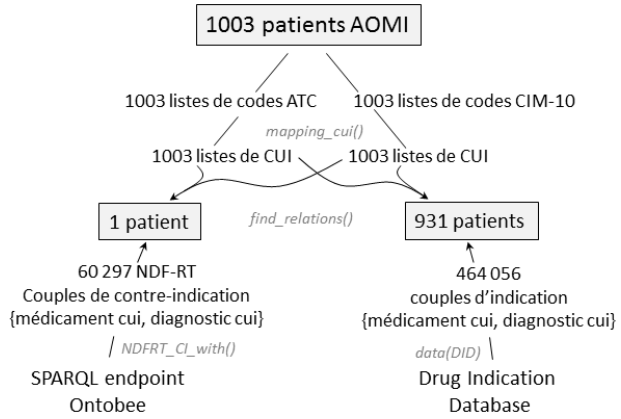

Figure 1 : Application de queryMed sur des données du SNDS. Les appels de fonction sont en gris.

La contre-indication était une prescription de vasoconstricteur, médicament visant à rétrécir les vaisseaux sanguins et contre-indiqué dans le contexte de l'AOMI car réduisant déjà la lumière des vaisseaux sanguins des membres inférieurs. Si l'interprétation d'une présence de relation, en l'occurrence une contre-indication entre une maladie et un médicament, est relativement aisée, celle de l'absence de relation reste beaucoup plus délicate. Les 72 patients sans relation d'indication entre leur état de santé d'AOMI et de médicament, peuvent à la fois refléter un manque d'exhaustivité des connaissances utilisées, des données elles-mêmes parfois incomplètes, ou bien sûr un réel défaut de prescription.

Parallèlement, nous avons répondu à ces questions via un requêtage SPARQL de ces mêmes données et connaissances aux formats du Web Sémantique, dans l'environnement adapté FUSEKI. Cette application a pu montrer que le package permettait de reproduire les mêmes résultats.

5. Conclusion et discussion

L'utilisation d'ontologies biomédicales et des technologies du Web Sémantique permet d'élargir le champs des possibles dans la réutilisation des bases de données médico-administratives, notamment en pharmaco-épidémiologie et pharmaco-vigilance. Une telle approche nécessite cependant l'adoption de plusieurs technologies, le regroupement de connaissances dispersée sur le Web ainsi qu'un travail de compréhension des schémas de représentation de ces connaissances. queryMed offre une façon

simplifiée de poursuivre cette approche pour l'exploration et secondairement l'analyse statistique de ces données depuis R.

6. Bibliographie

[AYV 15] AYVAZ S. et al., « Toward a complete dataset of drug-drug interaction information from publicly available sources », *Journal of Biomedical Informatics*, vol. 55, 2015, p. 206–217.

[CAL 13] CALLAHAN A. et al., « Bio2RDF Release 2 : Improved Coverage, Interoperability and Provenance of Life Science Linked Data », *The Semantic Web : Semantics and Big Data*, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, mai 2013, p. 200–212.

[FER 13] FERREIRA J. D. et al., « On the usefulness of ontologies in epidemiology research and practice », *Journal of Epidemiology and Community Health*, vol. 67, n° 5, 2013, p. 385–388.

[JON 16] JONQUET C. et al., « SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique », *JFIM : Journées Francophones d'Informatique Médicale*, e-health pour tous, Genève, Switzerland, juin 2016.

[KUR 15] KURBATOVA N. et al., « ontoCAT : Ontology traversal and search », 2015, R package version 1.29.0.

[LEH 15] LEHMANN J. et al., « DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia », *Semantic Web*, vol. 6, 2015, p. 167-195.

[ONG 17] ONG E. et al., « Ontobee : A linked ontology data server to support ontology term dereferencing, linkage, query and integration », *Nucleic Acids Research*, vol. 45, n° D1, 2017, p. D347–D352.

[R C 17] R CORE TEAM, « R : A Language and Environment for Statistical Computing », R Foundation for Statistical Computing, Vienna, Austria, 2017.

[ROU 18] ROUX J., GRIMAUD O., LERAY E., « Use of state sequence analysis for care pathway analysis The example of multiple sclerosis », *Statistical Methods in Medical Research*, , 2018, SAGE Publications.

[SHA 17] SHARP M. E., « Toward a comprehensive drug ontology : extraction of drug-indication relations from diverse information sources », *Journal of Biomedical Semantics*, vol. 8, n° 1, 2017.

[STE 06] STEVENS R., BODENREIDER O., LUSSIER Y. A., « SEMANTIC WEBS FOR LIFE SCIENCES », *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, , 2006, p. 112–115.

[Van 13] VAN HAGE W. R. et al., « SPARQL : SPARQL client », 2013, R package version 1.16.

[WHE 11] WHETZEL P. L. et al., « BioPortal : enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications », *Nucleic Acids Research*, vol. 39, n° Web Server issue, 2011, p. W541–545.

[WIL 14] WILLIGHAGEN E., « Accessing biological data in R with semantic web technologies », http ://dx.doi.org/10.7287/peerj.preprints.185v3, 2014.