PEGASE: A Knowledge Graph for Search and Exploration in Pharmacovigilance Data

Carlos Bobed, Laura Douze, Sébastien Ferré, Romaric Marcilly

PEGASE: A Knowledge Graph for Search and Exploration in Pharmacovigilance Data*

Carlos Bobed[1], Laura Douze[2], Sébastien Ferré[1], and Romaric Marcilly[2]

[1] Univ Rennes, CNRS, IRISA
Campus de Beaulieu, 35042 Rennes, France
{carlos.bobed-lisbona,sebastien.ferre}@irisa.fr
[2] Univ. Lille, INSERM, CHU Lille, CIC-IT / Evalab 1403
Centre d'Investigation clinique, EA 2694, F-59000 Lille, France
{laura.douze,romaric.marcilly}@univ-lille.fr

Abstract. Pharmacovigilance is in charge of studying the adverse effects of pharmaceutical products. In this field, pharmacovigilance specialists experience several difficulties when searching and exploring their patient data despite the existence of standardized terminologies (MedDRA). In this paper, we present our approach to enhance the way pharmacovigilance specialists perform search and exploration on their data. First, we have developed a knowledge graph that relies on the OntoADR ontology to semantically enrich the MedDRA terminology with SNOMED CT concepts, and that includes anonymized patient data from FAERS. Second, we have chosen and applied a semantic search tool, Sparklis, according to the user requirements that we have identified in pharmacovigilance.

1 Introduction

The continuous research and advances in pharmacology improve significantly our life quality. However, despite being thoroughly tested before being released, all the possible side effects of the new drugs cannot be foreseen. Thus, along advances in pharmacology, we need methods to discover those adverse effects to improve the safety and efficacy of drugs. Pharmacovigilance is defined by the World Health Organization as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem". In this work, we are concerned with supporting pharmacovigilance specialists in the search and exploration of their database of patient cases, which is generally the first step in the process of detecting new adverse effects of drugs.

In this context, the usefulness of standardized vocabularies to unify the codification of the reports is evident. MedDRA (Medical Dictionary for Drug Regulatory Activities)[3] is the vocabulary recommended by the ICH for the electronic transmission of individual case safety reports [2] to code adverse drug reactions (ADRs). However, as pointed out by Bousquet et al. [3], "its main limitation comes from its standard terminological format, which restricts the possibility of accessing terms based on their semantics". To solve this problem, Bousquet et al. proposed OntoADR [3], an ontology which makes it possible to work with MedDRA terms according to their actual semantics.

* This research is supported by ANR project PEGASE (ANR-16-CE23-0011-08), project TIN2016-78011-C4-3-R (AEI/ FEDER, UE), and DGA/FEDER.

[3] MedDRA® is a registered trademark of IFPMA (Int. Fed. Pharm. Manufact. and Assoc.)

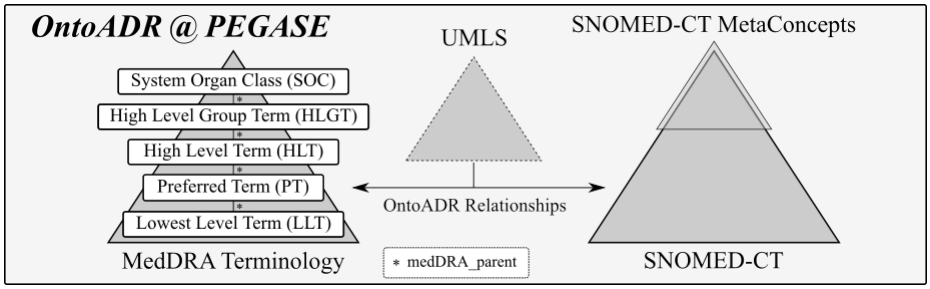

Fig. 1. Main modules of OntoADR within the PEGASE Knowledge Graph.

In this paper, we present the solution we have developed in the PEGASE project to improve the way pharmacovigilance specialists search for cases. First, we have built a knowledge graph based on OntoADR integrating different knowledge sources, which makes it possible to have all the relevant data easily accessible, providing the flexibility required to be extended under demand. Then, we have chosen and applied Sparklis [4], a query builder that eases the exploration and querying of any SPARQL endpoint, without requiring to master SPARQL itself. This choice was based on a requirement analysis conducted by ergonomists in the project. We are currently evaluating our proposal, along with other tools, in order to assess the benefits of our approach.

2 PEGASE Knowledge Graph

To build our knowledge graph, we have adapted OntoADR [3], extending it with SMQs (Standardised MedDRA Queries) [5] and anonymized patient data from FAERS dataset [1] to show how the integration capabilities of our knowledge graph can help pharmacovigilance specialists to ease their jobs.

OntoADR The core structure of the PEGASE Knowledge Graph can be seen in Figure 1. It currently contains 3,257,389 triples without taking into account FAERS data (with the patient data of three months, it grows to 28,125,629 triples). To model MedDRA, we have introduced the concept *MedDRATerm*, which has five different subconcepts corresponding to the five levels of their hierarchy (see Figure 1). However, to model the hierarchy relationship between terms, instead of using the *subclass* relationship (i.e., formal subsumption), we have introduced the property *medDRA_parent*. In this way, we can navigate the hierarchy without unexpected potential inferences.

To include SNOMED CT, we had to adapt its representation level. On the one hand, we had MedDRA terms, all of which were instances; on the other hand, we had SNOMED CT terms, all of which were concepts. To solve this mismatch, we materialized SNOMED CT concept hierarchy, and treated the concepts as instances[4]. This allowed us to introduce also different hierarchies to provide different navigation dimensions. In particular, we introduced a top-level hierarchy of SNOMED CT meta-concepts based on the semantic tags that SNOMED CT uses to further refine the concepts meaning. Note that this grouping cohabits with the subclass hierarchy of SNOMED CT concepts. This does not lead to inconsistencies as our knowledge graph is in RDFS, not in OWL.

[4] Abusing a little the language, we have flattened them in the RDF graph and allowed for meta-modeling, i.e., classes of SNOMED CT concepts.

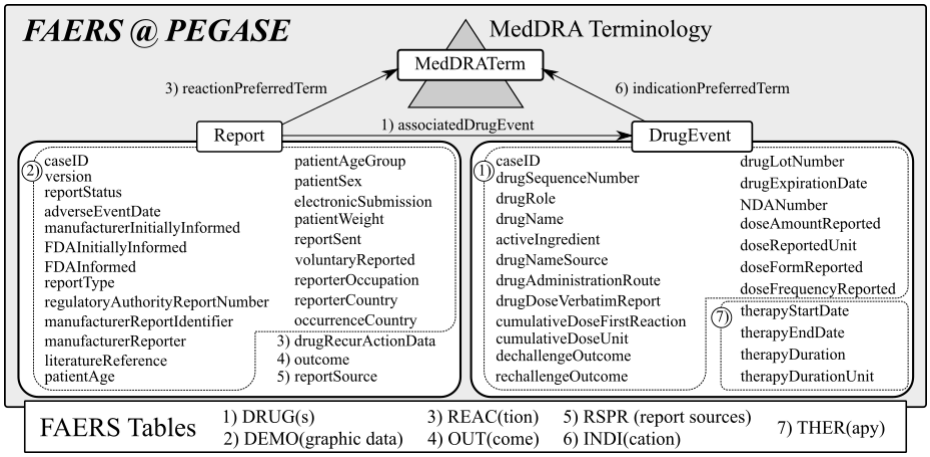

Fig. 2. FAERS integration in the knowledge graph. The companion numbers indicate the source of the data in the original FAERS tables.

OntoADR relationships between MedDRA terms and SNOMED CT concepts (see [3] for the complete list) were included as they are.

SMQs SMQs are "groupings of MedDRA terms, ordinarily at the Preferred Term (PT) level that relate to a defined medical condition or area of interest" [5]. In general, SMQs can be seen as disjunctions of terms which are used together in order to perform searches in a standardized way, although they can be grouped in more complex ways. We added each SMQ as a new node, related to the terms that it includes. The inclusion of SMQs is important because pharmacovigilants are used to work with them.

FAERS Data The patient data provided by FAERS is split in seven different big tables, which we have integrated as shown in the resulting model in Figure 2. That model was obtained after an evaluation round with the ergonomists in the project's team, where we brought the FAERS model closer to the pharmacovigilants cognitive process.

3 Sparklis on the PEGASE Knowledge Graph

Sparklis[5] is a query builder in natural language that allows people to explore and query SPARQL endpoints with all the power of SPARQL and without any knowledge of SPARQL [4]. It reconciles the expressivity of SPARQL 1.1 and the usability of point-and-click user interfaces. Sparklis requires little configuration to be applied to the PEGASE Knowledge Graph. It is enough to provide the URL of the SPARQL endpoint[6], and to choose property `rdfs:label` for the labelling of entities, classes, and properties.

Figure 3 shows a screenshot of Sparklis on PEGASE data, taken during the process of building a query[7]. The current query (at the top) select prefered terms (PT) in MedDRA whose finding site is (a subconcept of) "Skin and subcutaneous tissue structure", and

[5] `http://www.irisa.fr/LIS/ferre/sparklis/`

[6] The URL is not provided here due to restrictive licences on MedDRA and SNOMED.

[7] A screencast of the whole query building is available at `http://www.irisa.fr/LIS/common/documents/ekaw2018/#ExtraCase`.

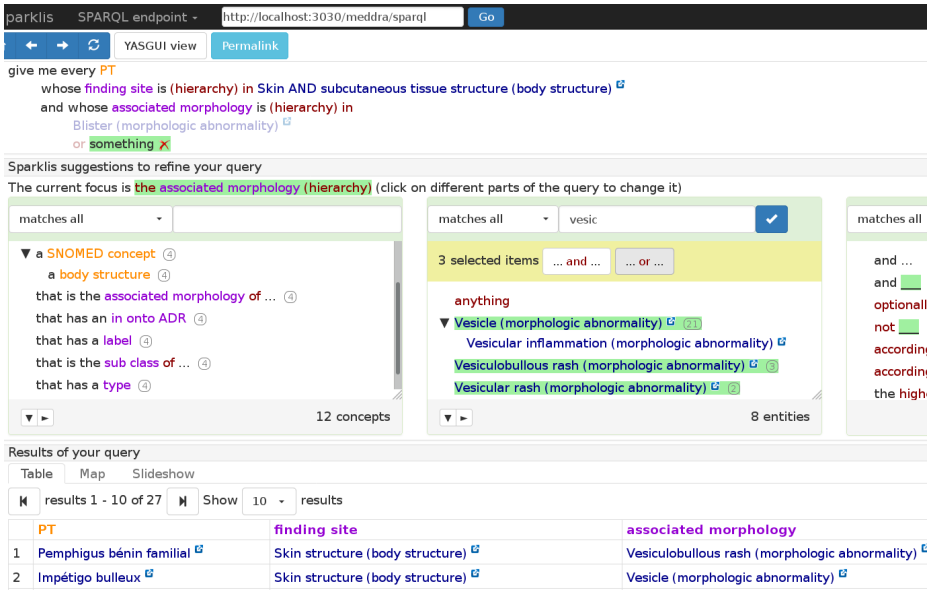

Fig. 3. Sparklis' screenshot showing a query under construction (top) on the PEGASE Knowledge Graph, suggestions to refine the query (middle), and query results (bottom).

whose associated morphology is (a subconcept of) various morphologic abnormalities. A first abnormality, "Blister" (dimmed font), has already been selected, and the user is in the process of selecting (at the center) a disjunction of three more abnormalities ("Vesicle", "Vesiculobullous rash", "Vesicular rash"). The keyword "vesic" was input at the top of the list of suggested terms in order to ease their retrieval among a long list of suggestions. The list of suggestions at the middle left contains classes and properties, i.e., types and relationships about the current focus (here, the focus is on the associated morphology of the selected preferred terms). The list of suggestions at the middle right contains query modifiers and operators (e.g., "and", "or", "number of"). The table of results of the current query is shown at each step (at the bottom). Here, it shows the selected preferred terms along with their finding sites and associated morphologies.

References

1. FDA's Adverse Event Reporting System (FAERS) Website. `https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm`, accessed: 9th July 2018.
2. ICH guideline E2B (R2), Electronic transmission of individual case safety reports, Final Version 2.3, Document Revision February, 2001.
3. Bousquet, C., Sadou, É., Souvignet, J., Jaulent, M.C., Declerck, G.: Formalizing MedDRA to support semantic reasoning on adverse drug reaction terms. Journal of Biomedical Informatics **49**, 282–291 (2014)
4. Ferré, S.: Sparklis: An expressive query builder for SPARQL endpoints with guidance in natural language. Semantic Web: Interoperability, Usability, Applicability **8**(3), 405–418 (2017)
5. ICH: Introductory Guide for Standardised MedDRA Queries (SMQs) Version 21.0, Document Revision March, 2018