Pharmaceutical knowledge base conception

Keywords: RDF, Graph database, data integration

Supervisors: Iovka.Boneva (@univ-lille.fr), Aurélien Lemay (@univ-lille.fr)

Location: Links project-team of Inria and CRIStAL, Haute Borne

Context:

It is estimated that every year, 10 000 people die in France due to bad usage of medical drugs. Reasons are diverse, ranging from error of prescription, dosage mistakes, drugs incompatible together or with a specific condition (illness or pregnancy for instance). Data that could prevent those mistakes usually exist and are freely available on the web, but they are scattered among hundreds of web sites. Furthermore, these websites can have very different presentations, they can also have different points of view which leads to inconsistency. Data can be stored in different formats that together with actual data formats (*rdf*, *xml)* includes raw text, *html*, *pdf* or *excel* spreadsheets for instance. The consequence is that it may be quite hard for a healthcare professional to find the relevant data.

This project is a partnership with **POSOS,** a new Company that proposes to solve this problem by creating a knowledge database that centralizes all relevant drug-related data, and a natural language system that allows an easy access.

Tasks:

The project focuses around the conception of the database in RDF. In particular, this includes the following subtasks:

- **Study of pharmaceutical websites**: in particular, study the heterogeneity of data encountered and their various formats. The objective is also to get familiar with the concept and the vocabulary used in pharmacology, and more general, in medicine.
- **The conception of the database:** we will use the ShEx format, developed by the Links Team, to describe the schema of the knowledge graph. The database also has to link each data to its source, in particular to deal with non-consistent data.
- **Data Integration:** the final phase consists in getting the actual data from the various websites. This means writing scripts that scrap various websites. More generally, this should be accompanied by a reflection on how to have a proper data integration framework with formalized tools based around the ShEx format. This should also takes into account the dynamic aspect of data, as websites evolve constantly over time.

Bibliography :

- **Ordoscopie** (www.ordoscopie.fr) *A website that references French pharmaceutical websites (under “boite à outils”, password :* pharcli*)*
- **LODD** (Linking Open Drug Data) (www.w3.org/wiki/HCLSIG/LODD) *a W3C initiative that aims at linking RDF graphs on drug data.*
- **Semantics and Validation of Shapes Schemas for RDF**. In ISWC2017 - 16th International semantic web conference. I. Boneva, J. Labra Gayo and E. Prud'hommeaux. *A presentation of the ShEx format*
- **RDF** : https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/Overview.html
- ***Data integration: A Theoretical Perspective.*** *M. Lenzerini. In the PODS 2002 Conference.*