UNIVERSITÉ DE RENNES 1

Direction de la Recherche et de l'Innovation

Appel à projets - Année 2023

Contrats doctoraux «Ordinaires»

ECOLE DOCTORALE : MATISSE **POLE DE RECHERCHE : INFO**

Unité de recherche d'accueil

Unité de Recherche : IRISA

Equipe : DYLISS

Encadrement

Directeur de thèse : Olivier DAMERON **Unité de recherche** : IRISA

Employeur/Établissement d'origine : Université de Rennes

HDR : • **oui** • ~~non~~ **Nb de thèses en cours** : 1

Mail : olivier.dameron@irisa.fr **Tél**. :

Co-directeur de thèse : **Unité de recherche** :

Employeur/Établissement d'origine :

HDR : • oui • non **Nb de thèses en cours** :

Mail : **Tél.** :

Sujet

Titre : Décomposition de requêtes SPARQL en modules de haut niveaux

Mots clés : Data science, bioinformatics, Semantic Web, SPARQL

Cotutelle internationale (oui/non) : **Si oui, préciser l'établissement et le pays de rattachement** :

Descriptif (15-30 lignes) :
Contexte socioéconomique et scientifique ; Les hypothèses et questions posées ; Les grandes étapes de la thèse et démarche ; Approches méthodologiques et techniques envisagées

Les sciences de la vie sont à la fois intrinsèquement compliquées à cause du grand nombre d'éléments différents qui entrent en jeu, et complexes à cause de l'interdépendance forte de ces éléments [1,2]. Pour étudier ces éléments, il y a 1645 bases de données de référence [3] qu'il faut pouvoir combiner [4]. Elles sont souvent adossées à des ontologies, qui formalisent les connaissances symboliques nécessaires à leur interprétation [5].
Les technologies du Web Sémantique (RDF pour la description des données, RDFS et OWL pour la formalisation des connaissances et SPARQL pour leur interrogation) offrent un cadre unifié aux problèmes d'intégration et d'interrogation qui sous-tendent les analyses [6]. Aujourd'hui, la plupart des bases de données et de connaissances en sciences de la vie sont disponibles dans ces formats et accessibles sous forme de SPARQL endpoint [7,8], et constituent la partie la plus dense et la plus inter-connectée des Linked Open Data [9]. Pourtant, (comme pour les LOD en général), elles demeurent singulièrement sous-exploitées. **La complexité des schémas de données de chacune de ces bases, qui reflète la complexité du domaine et est la garantie de la finesse des analyses, est également le principal facteur limitant leur appropriation par les non-informaticiens, qui constituent pourtant l'immense majorité des utilisateurs potentiels.** Deux moyens couramment utilisés pour dépasser cette difficulté sont la fourniture de collections d'exemples de requêtes, et de schémas décrivant les types d'entités leurs relations présentes dans les bases, mais ils se révèlent insuffisants.

Nous faisons l'hypothèse principale qu'**il est possible d'établir une représentation des données exposées dans un SPARQL endpoint sous forme de modules de haut niveau**. Chaque module correspond à une partie connexe du schéma de données. Les modules et leurs relations présentent le double avantage (1) de fournir une description fonctionnelle de haut niveau cognitif de la base de données, ancrée sur la représentation structurelle de ces données, et (2) de permettre la composition de requêtes en combinant des modules afin de générer automatiquement le code SPARQL correspondant.
Nous faisons l'hypothèse secondaire qu'**il est possible de générer une décomposition en modules d'un SPARQL endpoint (semi-)automatiquement à partir d'un ensemble de requêtes sur cet endpoint.**

La première étape de la thèse consistera à formaliser la notion de module ainsi que la façon dont on peut les assembler.

La seconde étape consistera à proposer une méthode de génération de modules à partir d'un ensemble de requêtes et à définir les critères d'optimalité (on cherche une décomposition en un nombre minimal de modules qui permette de couvrir tout le schéma de données et de composer toutes les requêtes ; cette décomposition n'étant pas nécessairement unique).
La troisième étape consistera à valider et évaluer la méthode précédente sur des bases de données de référence comme neXtProt, UniProt ou Reactome.

La validation et l'évaluation s'appuieront notamment sur la base neXtProt [10] qui est associée à 777 requêtes SPARQL validées et annotées à la main par des tags (148 sont utilisées dans le tutorial, et les autres servent au contrôle d'intégrité).

[1] Carol J. Bult. From information to understanding : the role of model organism databases in comparative and functional genomics. Animal Genetics, 37(suppl. 1) :28–40, 2006.
[2] Olivier Bodenreider and Robert Stevens. Bio-ontologies : current trends and future directions. Briefings in Bioinformatics, 7(3) :256–274, 2006.
[3] Daniel J Rigden and Xosé M Fernández. The 2022 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic acids research, 50(D1) :D1–D10, 2022.
[4] Nicola Cannata, Emanuela Merelli, and Russ B. Altman. Time to organize the bioinformatics resourceome. PLoS Computational Biology, 1(7) :0531–0533, 2005.
[5] Judith A. Blake and Carol J. Bult. Beyond the data deluge : Data integration and bio-ontologies. Journal of Biomedical Informatics, 39(3) :314–320, 2006.
[6] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. Scientific American,
284(5) :34–43, 2001.
[7] Susie Stephens, David LaVigna, Mike DiLascio, and Joanne Luciano. Aggregation of bioinformatics data using semantic web technology. Journal of Web Semantics, 4(3), 2006.
[8] Nicola Cannata, Michael Schröder, Roberto Marangoni, and Paolo Romano. A semantic web for bioinformatics : goals, tools, systems, applications. BMC bioinformatics, 9 Suppl 4 :S1, 2008.
[9] Christian Bizer, Tom Heath, and Tim Berners Lee. Linked data–the story so far. International Journal on Semantic Web and Information Systems, 5(3) :1–22, 2009.
[10] Monique Zahn-Zabal, Pierre-André Michel, Alain Gateau, Frédéric Nikitin, Mathieu Schaeffer, Estelle Audot, Pascale Gaudet, Paula D Duek, Daniel Teixeira, Valentine Rech de Laval, Kasun Samarasinghe, Amos Bairoch, and Lydie Lane. The nextprot knowledgebase in 2020 : data, tools and usability improvements. Nucleic acids research, 48(D1) :D328–D334, 2020.