

# Techniques coopératives pour l'exploitation des bases de connaissances et passage à l'échelle

**Laboratoire :** Laboratoire d'Informatique et d'Automatique pour les Systèmes<sup>1</sup>

**Localisation :** Poitiers (locaux de l'ISAE-ENSMA)

**Financement :** Bourse Ministérielle avec possibilité de vacances à l'ISAE-ENSMA et à l'Université de Poitiers

**Équipe encadrante :** Stéphane JEAN, Brice CHARDIN, Allet HADJALI, Mickaël BARON

**Date limite de dépôt des candidatures :** 7 mai 2019

**Mots-clés :** bases de connaissances, treillis, relaxation, RDF, SPARQL

## Contexte

Avec l'émergence et la multiplication des applications du Web sémantique, de nombreuses bases de connaissances, à la fois récentes, volumineuse et potentiellement incertaines deviennent disponibles. Ces bases de connaissances contiennent des entités nommées et des faits sur ces entités, mais aussi les classes sémantiques de ces entités et leurs liens mutuels. De plus, plusieurs bases de connaissance peuvent être interconnectées au niveau de leurs entités, formant ainsi le noyau du Web des données liées (ou ouvertes). Parmi les bases de connaissance les plus connues, citons [Yago](#), [DBPedia](#), [Nell](#), [DeepDive](#), [Google's Knowledge Vault](#) et [Freebase](#).

Au LIAS, nous cherchons à proposer des techniques facilitant la gestion et l'exploitation des bases de connaissances, en particulier pour aider l'utilisateur lorsque sa requête ne lui retourne aucun résultat. Les techniques développées consistent essentiellement à identifier les causes d'échec de cette requête. Ces techniques reposent sur l'exploration du treillis formé par l'ensemble des sous-requêtes. Ce treillis étant d'une taille exponentielle par rapport à la taille de la requête originale, l'exploration s'appuie sur des heuristiques et des propriétés de monotonie permettant d'élaguer l'espace de recherche.

## Sujet de la thèse

Jusqu'à présent, nous n'avons considéré que le problème des réponses vides. Cependant, d'autres types de résultats peuvent être considérés comme insatisfaisants par l'utilisateur, menant à de nouvelles notions d'échec telles que :

- *réponses insuffisantes* : l'utilisateur n'obtient qu'un nombre faible de résultats par rapport à ses attentes;
- *réponses pléthoriques* : l'utilisateur obtient un nombre trop important de résultats qui l'empêche d'en extraire l'information qui lui est pertinente;
- *absence d'un résultat* : l'utilisateur s'attendait à obtenir un résultat qui n'apparaît pas dans la réponse à sa requête.

Pour répondre à ces trois problèmes, peu de travaux se sont intéressés à identifier les causes d'échec de la requête utilisateur. Aussi, un premier objectif de la thèse sera d'étudier comment adapter ou étendre les approches développées pour le problème des réponses vides à ces nouveaux contextes. Le défi principal est que, dans le contexte du problème des réponses vides, la monotonie de la propriété considérée

1. <https://www.lias-lab.fr>



(l'échec) permet d'élaguer progressivement l'espace de recherche. Cette monotonie n'est plus forcément garantie pour ces nouveaux problèmes. Aussi un travail théorique devra être mené pour identifier sous quelles conditions cette propriété est maintenue et, si ce n'est pas le cas, quelles solutions alternatives peuvent être proposées pour explorer efficacement l'espace des sous-requêtes.

D'autre part, les expérimentations que nous avons menées pour le problème des réponses vides ont montré que nos approches proposent des temps de réponse raisonnables lorsqu'elles sont utilisées sur des bases de connaissance contenant des millions de faits. Dans le contexte des données massives, des volumes de données plus importants et des requêtes plus complexes peuvent être considérés. Pour répondre efficacement à ce besoin de passage à l'échelle, il devient nécessaire d'utiliser des solutions distribuées sur plusieurs nœuds de traitement. Ainsi, un second objectif de la thèse sera d'étudier comment les approches que nous avons proposées pourraient être parallélisées. Plus généralement, le travail mené dans cette thèse s'intéressera à l'ajout de techniques d'optimisation pour améliorer le temps de traitement des approches proposées.

### Publications associées

- Ibrahim DELLAL, Stéphane JEAN, Allel HADJALI, Brice CHARDIN, Mickael BARON, Query Answering over Uncertain RDF Knowledge Bases : Explain and Obviate Unsuccessful Query Results, Knowledge and Information Systems (KAIS 2019), 2019
- Ibrahim DELLAL, Stéphane JEAN, Allel HADJALI, Brice CHARDIN, Mickael BARON, On Addressing the Empty Answer Problem in Uncertain Knowledge Bases, Database and Expert Systems Applications (DEXA 2017), vol. 10438, LNCS, edited by Springer, 2017, pp. 120-129
- Géraud FOKOU, Stéphane JEAN, Allel HADJALI, Mickael BARON, RDF Query Relaxation Strategies Based on Failure Causes (Best Research Paper Award), 13th International Extended Semantic Web Conference, Heraklion, Greece (ESWC16), 2016, pp. 439-454

### Profil du candidat

Le candidat devra être titulaire d'un Master 2 ou d'un diplôme d'ingénieur et posséder des connaissances en algèbre, traitement de données, administration de bases de données et programmation. Un bon niveau en français et en anglais est également nécessaire.

Les dossiers de candidature sont à envoyer à [stephane.jean@ensma.fr](mailto:stephane.jean@ensma.fr), avec les documents suivants :

- Curriculum Vitae,
- Lettre de motivation,
- Notes de Master ou équivalent,
- Tout autre document jugé nécessaire par le candidat pour enrichir le dossier de candidature.

