COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

# THÈSE DE DOCTORAT

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
*Mathématiques et Sciences et Technologies du numérique de l'Information et de la Communication*
Spécialité : *Informatique*

Par
**Julien AIMONIER-DAVAT**

## Querying Online Decentralized Knowledge Graphs

**Thèse présentée et soutenue à Nantes, le 11 décembre 2023**
**Unité de recherche : Laboratoire des sciences du numérique à Nantes (LS2N)**

### Rapporteurs avant soutenance :

| | |
|---|---|
| Frédérique LAFOREST | Professeur des universités, INSA Lyon |
| Mathieu D'AQUIN | Professeur des universités, Université de Lorraine, Nancy |

### Composition du Jury :

| | | |
|---|---|---|
| Président : | Prénom NOM | Fonction et établissement d'exercice *(à préciser après la soutenance)* |
| Examinateurs : | Frédérique LAFOREST | Professeur des universités, INSA Lyon |
| | Mathieu D'AQUIN | Professeur des universités, Université de Lorraine, Nancy |
| | Arnaud SOULET | Professeur des universités, Université de Tours |
| | Luis CALARRAGE | Chargé de recherches, INRIA Rennes |
| | Mounira HARAZALLAH | Maître de conférences des universités HDR, Nantes Université |
| Dir. de thèse : | Hala SKAF-MOLLI | Professeur des universités, Nantes Université |
| Co-dir. de thèse : | Pascal MOLLI | Professeur des universités, Nantes Université |

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

**Titre :** Comment interroger un graphe de connaissance décentralisé en ligne ?

**Mot clés :** Web sémantique, serveurs SPARQL publics, préemption web, optimisation, moteur de requêtes fédérées

**Résumé :**

Dans l'état actuel du web sémantique, développer des applications intelligentes au dessus d'un graphe de connaissances décentralisé interrogeable en ligne reste un rêve lointain. D'une part, les serveurs SPARQL publics font face à de grave problèmes d'accessibilité. D'autre part, les moteurs de requêtes fédérés qui permettent aux utilisateurs d'interroger de manière transparente plusieurs serveurs SPARQL ne passent pas à l'échelle. Ainsi, des milliards de triples RDF sont disponibles mais non accessibles. Différentes solutions ont été proposées pour résoudre ces problèmes mais les solutions existantes présentent toutes des problèmes de performance importants. Dans cette thèse, nous proposons des solutions pour remédier à ces limitations. Nous commençons par étendre le modèle de la préemption du Web avec de nouveaux opérateurs. Ceci a permis d'améliorer significativement les performances de certaines classes de requêtes parmi les plus utilisées. Ensuite, nous proposons un nouvel algorithme d'optimisation pour les requêtes SPARQL conjonctives avec des filtres et des property paths. Enfin, nous introduisons un nouveau moteur de requêtes fédérées qui permet de réduire le nombre de sources à contacter pour exécuter une requête.

**Title:** Querying Online Decentralized Knowledge Graphs

**Keywords:** Semantic Web, Public SPARQL Servers, Web Preemption, Join Ordering, Federated Query Engine

**Abstract:**

In the current state of the Semantic Web, developing intelligent applications on top of a live queryable decentralized knowledge graph remains a distant dream. On the one hand, public SPARQL endpoints face serious availability problems. On the other hand, federated query engines that allow users to query multiple endpoints as a single one do not scale when the number of sources increases. As a result, billions of RDF triples are available but not accessible. Various solutions have been proposed to solve these problems, but existing solutions all present significant performance issues. In this thesis, we propose solutions to overcome these limitations. We start by extending the Web preemption model with new preemptable operators. In doing so, we significantly improve the performance of frequently used classes of queries. Next, we propose a new optimization algorithm for conjunctive SPARQL queries with filters and property paths. Finally, we introduce a new federated query engine that reduces the number of sources to be contacted to execute a query.