

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE NO. 641

*Mathématiques et Sciences et Technologies du numérique,
de l'Information et de la Communication*

Spécialité : *Informatique*

Par

Thi Hoang Thi PHAM

**From Querying to Answering with Online Knowledge Graphs
under Real-World Constraints**

Thèse présentée et soutenue à Nantes, le 22 octobre 2026

Unité de recherche : LS2N - Laboratoire des Sciences du Numérique de Nantes

Rapporteurs avant soutenance :

Fabien GANDON Directeur de recherche, Inria, Université Côte d'Azur
Arnaud SOULET Professeur des universités, Université de Tours

Composition du Jury :

Présidente :

Examineurs : Fabien GANDON Directeur de recherche, Inria, Université Côte d'Azur
Arnaud SOULET Professeur des universités, Université de Tours
Peggy CELLIER Maître de conférences HDR, INSA Rennes, IRISA
Richard DUFOUR Professeur des universités, Nantes Université

Dir. de thèse : Pascal MOLLI Professeur des universités, Nantes Université

Co-dir. de thèse : Hala SKAF-MOLLI Professeure des universités, Nantes Université

Titre : De l'interrogation à la réponse avec des graphes de connaissances en ligne sous contraintes réelles

Mots clés : points d'accès SPARQL, count-distinct, requêtes de continuation SPARQL, réponse aux questions hybride, complétude

Résumé : Les grands graphes de connaissances publics exposent des données RDF au moyen de points d'accès SPARQL, mais les quotas et les limites de temps d'exécution peuvent empêcher ces points d'accès de retourner le résultat exact d'une requête count-distinct ou le résultat complet d'une autre requête. Un graphe de connaissances en ligne peut également ne pas contenir certains faits requis pour répondre à une question en langage naturel. Cette thèse traite ces trois cas d'échec avec des méthodes, des types de réponse et des hypothèses distincts. Elle propose CRAWD, un estimateur fondé sur l'échantillonnage pour les requêtes

SPARQL count-distinct sur des points d'accès SPARQL publics qui permettent l'échantillonnage requis lorsque l'énumération exacte dépasse le budget disponible ; PASSAGE, qui utilise des requêtes de continuation SPARQL pour retrouver le résultat complet d'une requête du cœur de SPARQL sur un graphe RDF fixe, sous les hypothèses d'énumération déterministe des motifs de triplets et de progrès ; et KARMA, qui utilise le schéma projeté autour des entités liées pour décider quand tenter une évaluation SPARQL et recherche des preuves dans le texte source lié lorsque la couverture est insuffisante ou que SPARQL ne retourne aucun résultat.

Title: From Querying to Answering with Online Knowledge Graphs under Real-World Constraints

Keywords: public SPARQL endpoints, count-distinct, SPARQL continuation queries, hybrid question answering, completeness

Abstract: Large public Knowledge Graphs expose RDF data through SPARQL endpoints, but quotas and timeouts can prevent endpoints from returning exact count-distinct values or complete query results. An online Knowledge Graph may also lack some facts required to answer a question in natural language. This thesis addresses these three failure cases with distinct methods, answer types, and assumptions. It contributes CRAWD, a sampling-based estimator for count-distinct SPARQL queries over public SPARQL end-

points that support sampling when exact enumeration exceeds the available budget; PASSAGE, which uses SPARQL continuation queries to recover complete query results for core SPARQL over a fixed RDF graph under assumptions of deterministic triple pattern enumeration and progress; and KARMA, which uses the projected schema around linked entities to decide when to attempt SPARQL and retrieves evidence from linked source text when schema coverage fails or SPARQL returns no bindings.