



HAL
open science

Scalable Reasoning on Document Stores via Instance-Aware Query Rewriting

Olivier Rodriguez, Federico Ulliana, Marie-Laure Mugnier

► **To cite this version:**

Olivier Rodriguez, Federico Ulliana, Marie-Laure Mugnier. Scalable Reasoning on Document Stores via Instance-Aware Query Rewriting. BDA 2023 - 39e Conférence sur la Gestion de Données – Principes, Technologies et Applications, Oct 2023, Montpellier, France. lirmm-04305822

HAL Id: lirmm-04305822

<https://hal-lirmm.ccsd.cnrs.fr/lirmm-04305822>

Submitted on 24 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0
International License

Scalable Reasoning on Document Stores via Instance-Aware Query Rewriting

Olivier Rodriguez
LIRMM, Inria, Univ. Montpellier,
CNRS
Montpellier, France
olivier.rodriguez@inria.fr

Federico Ulliana
LIRMM, Inria, Univ. Montpellier,
CNRS
Montpellier, France
federico.ulliana@inria.fr

Marie-Laure Mugnier
LIRMM, Inria, Univ. Montpellier,
CNRS
Montpellier, France
marie-laure.mugnier@inria.fr

ABSTRACT

Data trees, typically encoded in JSON, are ubiquitous in data-driven applications. This ubiquity makes urgent the development of novel techniques for querying heterogeneous JSON data in a flexible manner. We propose a rule language for JSON, called constrained tree-rules, whose purpose is to provide a high-level unified view of heterogeneous JSON data and infer implicit information. As reasoning with constrained tree-rules is undecidable, we identify a relevant subset featuring tractable query answering, for which we design an automata-based query rewriting algorithm. Our approach consists of leveraging NoSQL document stores by means of a novel

instance-aware query-rewriting technique. We present an extensive experimental analysis on large collections of several million JSON records. Our results show the importance of instance-aware rewriting as well as the efficiency and scalability of our approach.

ACKNOWLEDGMENTS

This work was financially supported by the ANR project CQFD (ANR-18-CE23-0003).

© 2023, Copyright is with the authors. Published in the Proceedings of the BDA 2023 Conference (October 23-26, 2023, Montpellier, France). Distribution of this article is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.

© 2023, Droits restant aux auteurs. Publié dans les actes de la conférence BDA 2023 (23-26 octobre 2023, Montpellier, France). Redistribution de cet article autorisée selon les termes de la licence Creative Commons CC-by-nc-nd 4.0.