

Towards Practicable Algorithms for Rewriting Graph Queries beyond DL-Lite (Extended Abstract)

Bianca Löhnert¹, Nikolaus Augsten¹, Cem Okulmus² and Magdalena Ortiz³

¹Paris-Lodron University of Salzburg, Austria ²Umeå University, Sweden ³TU Wien, Austria


Abstract


Despite the many advantages that ontology-based data access (OBDA) has brought to a range of application domains, state-of-the-art OBDA systems still do not support popular graph database management systems such as Neo4j. Algorithms for query rewriting focus on languages like conjunctive queries and their unions, which are fragments of first-order logic and were developed for relational data. Such query languages are poorly suited for querying graph data. Moreover, they also limit the expressiveness of the ontology languages that admit rewritings, restricting them to those where the data complexity of reasoning is not higher than it is in first-order logic. In this paper, we propose a technique for rewriting a family of navigational queries for a suitably restricted fragment of \mathcal{ELHI} that extends DL-Lite and that is NL-complete in data complexity. We implemented a proof-of-concept prototype that rewrites into Cypher queries, and tested it on a real-world cognitive neuroscience use case with promising results.

1. Introduction


The ontology-based data access (OBDA) paradigm has seen successful adoption and brought significant advantages to a range of applications [1], but so far it remains limited to relational database management systems (RDBMS). Recent years have seen huge adoption of *graph databases* and bringing the OBDA paradigm could open up a plethora of novel opportunities.


The central problem in OBDA is *ontology-mediated query answering* (OMQA), where a query is to be evaluated over the consequences of the given dataset together with the knowledge in the ontology. The predominant technique for OMQA is *query rewriting*, where an input query is transformed to incorporate the ontological knowledge so that the rewritten query can be directly evaluated over any input dataset using standard technologies. Until now, this has meant that the target of query rewriting have been variants of *conjunctive queries* (CQs), expressible in SQL and supported in standard RDBMs [2], and the ontology languages have been limited to those whose inference problem is not harder in data complexity than SQL evaluation. The DL-Lite family of description logics (DLs), designed to fulfil this requirement, remains the ontology language of choice for OMQA [2]. But with graph query languages, breaking through this barrier is possible. Their *navigational features* allow queries to traverse paths of arbitrary length that comply with some *regular path expression*, which results in a higher data complexity than that of SQL. Standard graph query languages like *conjunctive two-way regular path queries* (C2RPQs), the most popular navigational extension of CQs, and GQL [3], a new

 DL 2024: 37th International Workshop on Description Logics, June 18–21, 2024, Bergen, Norway

 bianca.loehnert@plus.ac.at (B. Löhnert); nikolaus.augsten@plus.ac.at (N. Augsten); okulmus@cs.umu.se (C. Okulmus); magdalena.ortiz@tuwien.ac.at (M. Ortiz)

 0000-0002-3036-6201 (N. Augsten); 0000-0002-7742-0439 (C. Okulmus); 0000-0002-2344-9658 (M. Ortiz)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

standard for extending C2RPQs, are NL complete in data complexity (under the so-called *walk semantics*). This means that we are not bound to DL-Lite and can attempt query rewriting for richer ontology languages, provided that they have NL data complexity.

Contributions This extended abstract summarizes the following results, described in [4].

- We prove the existence of C2RPQs that cannot be rewritten into a union of C2RPQs (UC2RPQs) which captures all consequences from the ontology, even if the ontology is restricted to one axiom $\exists r \sqsubseteq \exists s$ for role names r, s . Motivated by this negative result, we define Navigational Conjunctive Queries (NCQs), a subset of C2RPQs (similar to [5, 6]) which admits such rewritings.
- We propose such a rewriting for OMQs that pair NCQs with ontologies in a language that we call $\mathcal{ELHI}_{\perp}^{lin}$, tailored to extend DL-Lite with some features of \mathcal{ELHI} while keeping the data complexity of standard reasoning in NL. We first show how to construct a graph structure whose paths witness all the relevant entailments of the input ontology, and then use this graph to rewrite atomic queries into C2RPQs. Then we combine this with the ideas behind the Clipper rewriting [7] to obtain a rewriting of NCQ mediated by $\mathcal{ELHI}_{\perp}^{lin}$ ontologies into UC2RPQs.
- We present a proof-of-concept prototype of our technique and use it to evaluate queries over data from a real-world use case in the domain of cognitive neuroscience.

2. Query Rewriting Algorithm for Navigational Queries

For our algorithm we consider $\mathcal{ELHI}_{\perp}^{lin}$ TBoxes, that have axioms of the form (NF1) $A \sqsubseteq B$, (NF2) $\exists r.A \sqsubseteq B$, (NF3) $A \sqsubseteq \exists r.B$, (NF4) $r \sqsubseteq s$, (NF5) $\exists r^{-}.\top \sqsubseteq B$, or (NF6) $A \sqsubseteq \exists r^{-}.\top$.

In *Navigational Conjunctive Queries* (NCQs), atoms take the form $(A_1 \cup \dots \cup A_n)(x)$ or $(\pi_1 \cup \dots \cup \pi_n)(x, y)$ or $(\pi_1 \cup \dots \cup \pi_n)^*(x, y)$, with A_i a concept name and each π_i a restricted *regular path expression* $\pi_i := r \mid r^{-} \mid r^* \mid (r^{-})^*$. As a first step towards our query rewriting algorithm, we introduce the so-called *concept dependency graph* for $\mathcal{ELHI}_{\perp}^{lin}$ TBoxes.

Concept Dependency Graph As the name suggests the concept dependency graph (CDG) depicts the dependencies of concepts of a given TBox. The nodes of the CDG represent concepts and for axioms in form of (NF1), e.g., $A \sqsubseteq B$ we add an edge pointing from B to A with the empty label ε . In case of an existential restriction on the left-hand side as in (NF2) or (NF5) we additionally label this edge with the role name. In order to make the construction of the CDG complete for TBox reasoning we extend it with additional edges and witnessing nodes for axioms with existential quantifiers on the right-hand side, i.e., (NF3) and (NF6). Having all these edges in place it is possible to check for TBox entailments. More precisely, it holds that $\mathcal{T} \models \exists r_1 \dots \exists r_k.A \sqsubseteq B$ if and only if there is a path from B to A with a sequence of edge labels s_1, \dots, s_k (possibly interleaved with ε -labeled edges) such that $s_i \sqsubseteq_{\mathcal{T}}^* r_i$ for $1 \leq i \leq k$. Note that, in particular, this means that $\mathcal{T} \models A \sqsubseteq B$ iff there is a path from B to A with ε labels only. We illustrate these properties of the CDG in the example below. For a formal definition of concept dependency graphs see [4].

Example 1. For the TBox $\mathcal{T} = \{A_2 \sqsubseteq A_1, \exists r.B_1 \sqsubseteq A_1, \exists r_3.B_1 \sqsubseteq B_3, A_3 \sqsubseteq A_2, \exists r_1.B_2 \sqsubseteq B_1, \exists r_2^{-}.\top \sqsubseteq A_3, s \sqsubseteq r_2, \exists r_2.B_3 \sqsubseteq B_2, B_1 \sqsubseteq \exists r_2.B_3\}$ the CDG $G_{\mathcal{T}}$ is as in Figure 1. The presence of paths in the CDG corresponds to concept inclusions that hold in \mathcal{T} . We illustrate this on an ABox $\mathcal{A} = \{r_1(a_1, a_2), r_2(a_2, a_3), r_3(a_3, a_4), r_1(a_4, a_5), B_2(a_5)\}$. Since there is the path

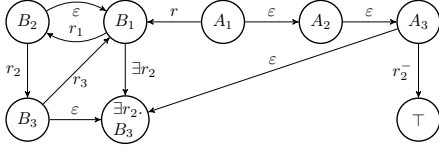


Figure 1: CDG from Example 1.

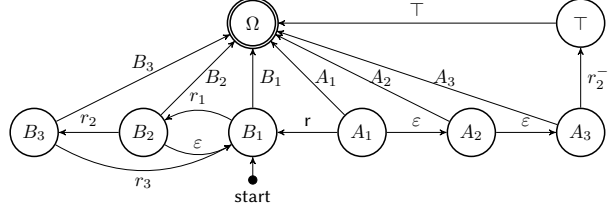


Figure 2: NFA from Example 2.

$B_1 r_1 B_2 r_2 B_3 r_3 B_1 r_1 B_2$ in the CDG of \mathcal{T} it follows for a_1 that $\mathcal{T}, \mathcal{A} \models B_1(a_1)$. However, as we mentioned before, we can use the CDG to check for TBox entailments, i.e, from the path $B_1 r_1 B_2 r_2 B_3 r_3 B_1 r_1 B_2$ we can infer that $\mathcal{T} \models \exists r_1 \exists r_2 \exists r_3 \exists r_1. B_2 \sqsubseteq B_1$. The construction of the CDG is complete, in the sense that for every $\mathcal{ELHI}_{\perp}^{lin}$ entailment from \mathcal{T} with a concept name on the right-hand side there exists a corresponding path in the CDG of \mathcal{T} .

Extracting regular path expressions. Given a CDG $G_{\mathcal{T}}$ and a concept B in \mathcal{T} , we construct an NFA $G_{\mathcal{T}}^A(B)$ that accepts the propagating paths in $G_{\mathcal{T}}$ which start at node B and end in another concept while passing over only roles or ϵ , as illustrated in Example 2. We then use an off-the-shelf technique to transform this NFA into a regular expression, and finally we eliminate all occurrences of ϵ from this regular expression.

Example 2. For the CDG from Example 1, we can see its B_1 -path-generating NFA in Figure 2.

The extracted RPE is: $\left(\overbrace{((r_1 \cup (r_1 r_2 r_3))^* B_1)}^{\text{paths ending in } B_1} \cup \overbrace{(r_1^+ (r_2 r_3 r_1^+)^* B_2)}^{\text{paths ending in } B_2} \cup \overbrace{(r_1^+ r_2 (r_3 r_1^+ r_2)^* B_3)}^{\text{paths ending in } B_3} \right)$.

Rewriting Atomic Queries Into Regular Path Expressions We replace each A_i in a given input of the form $(A_1 \cup \dots \cup A_n)(x)$ with the regular expression described above. Then we replace each role r in the regular path expression with a union over all roles s such that $s \sqsubseteq_{\mathcal{T}}^* r$.

Rewriting NCQs into UC2RPQs Given a NCQ $q(\vec{x})$ and an $\mathcal{ELHI}_{\perp}^{lin}$ TBox \mathcal{T} , we exhaustively apply the so-called *clipping function*, inspired by [7]. It relies on a well-known property of \mathcal{ELHI} : for each \mathcal{T} and \mathcal{A} there is a universal, forest-shaped model \mathcal{I}^u that can be used for answering all C2RPQs. Intuitively, we consider each possible set Y of variables that may be mapped to the same ‘anonymous’ object d of maximal depth in a tree structure of \mathcal{I}^u . We know that each such d is introduced to satisfy an existential axiom $A \sqsubseteq \exists r. B$, that is, it has a parent d_p that satisfies A , and d is created as an r -child of d_p that is B . Exploiting this, we can ‘remove’ from the query the parts that are guaranteed to be locally satisfied using d_p ’s r -child d . This idea has been applied to variants of (U)CQs [7], where each application must only remove the query atoms that are locally satisfied. But in navigational queries we must consider also how path expressions can be partially satisfied, and modify the expressions in the atoms accordingly. E.g., given $q(x) = (r^* \cup s^*)(x, y), B(y)$ and $\mathcal{T} = \{A \sqsubseteq \exists r. B\}$, a possible application of clipping returns $q'(x) = r^*(x, y), A(y)$. Note that $r^* \cup s^*$ is replaced by r^* , reflecting the fact that an r^* path to an A can always be extended to an r^* path to a B in the models of \mathcal{T} , but this does not apply to the s^* paths. We show in [4] that for every query mapping there is an application of

Table 1

Experimental results of the algorithm. $\#V$ is the number of C2RPQs, $\#\alpha$ the number of atoms and $\#C$ the number of distinct concepts in the output, \varnothing_{\cup} is the average and \max_{\cup} the maximal number of expressions in unions inside the C2RPQs, t_{rew} and t_{eval} the times for rewriting and evaluation in seconds. We indicate the two ontologies by O_1 and O_2 . Timeout was 1000 seconds.

	$\#V$	$\#\alpha$	$\#C^{O_1}$	$\#C^{O_2}$	$\varnothing_{\cup}^{O_1}$	$\varnothing_{\cup}^{O_2}$	$\max_{\cup}^{O_1}$	$\max_{\cup}^{O_2}$	$t_{\text{rew}}^{O_1}$	$t_{\text{rew}}^{O_2}$	$t_{\text{eval}}^{O_1}$	$t_{\text{eval}}^{O_2}$
Q1	60	313	934 957	987 1007	995 1062	17.78 32.76	904.13 800.25					
Q2	23	123	951 972	995 1020	1036 1062	7.44 8.77	253.60 245.77					
Q3	1	3	3 3	2 2	2 2	0.08 0.08	0.23 0.71					
Q4	1	3	3 3	2 2	2 2	0.07 0.07	0.11 0.03					
Q5	1	3	12 12	11 11	11 11	0.07 0.06	2.08 1.94					
Q6	297	1891	1018 1042	1063 1085	1169 1178	131.55 132.31	>1000 >1000					
Q7	77	509	922 932	988 1002	994 1009	32.24 54.20	>1000 >1000					
Q8	1	4	2 2	0 0	0 0	0.2 0.2	0.08 0.08					

clipping that results in a rewritten query whose mappings have a strictly lower depth (as at least one variable is now mapped to d_p instead of its child d), but remain unchanged otherwise.

By clipping exhaustively, iterating over all sets Y of variables in q and all axioms in \mathcal{T} of the form $A \sqsubseteq \exists r.B$ and $A \sqsubseteq \exists r^-. \top$, and taking each result as a disjunct in a union of NCQs, we obtain a query that has the same answers as the original NCQ, but whose variables are mapped to ABox individuals only. We can then apply the rewriting of atomic queries into regular path expressions described above, and obtain a sound and complete rewriting of NCQs mediated by $\mathcal{ELHI}_{\perp}^{\text{lin}}$ ontologies into UC2RPQs.

3. Implementation and Experiments

We implemented a proof-of-concept prototype [8] that, given an $\mathcal{ELHI}_{\perp}^{\text{lin}}$ TBox (in OWL syntax), rewrites NCQs into UC2RPQs and translates them into Cypher. We execute the experiments on a machine running Ubuntu 22.04.4 with an Intel Core i7-6700HQ CPU clocked at 2.60 GHz and 8 GB RAM; with Neo4j 5.18.1 running on it. As TBox (OWL ontology) we use the Cognitive Task Ontology (COGITO) [9], which includes about 9000 axioms describing 4700 concepts used for annotating experimental data [10]. For example, the axiom $\text{ReadingTask} \sqsubseteq (\exists \text{has.Read} \sqcap \exists \text{has.Language-item})$ defines a reading task by referring to the Hierarchical Event Descriptors (HEDs) Read and Language-item [11]. COGITO includes axioms that are not in $\mathcal{ELHI}_{\perp}^{\text{lin}}$ since they use conjunction on the left side (e.g., the converse of the ReadingTask axiom above). Hence, to make our experimental evaluation more faithful to COGITO, we consider two ontologies: (1) in O_1 we drop the axioms with conjunction on the left-hand side, and (2) in O_2 we replace these conjunctions by disjunctions. We are currently working on inferring the exact answers over COGITO from these over- and under-approximations. We chose a real-world dataset from the domain of cognitive neuroscience [12], stored in a Neo4j database and consisting of 15 512 nodes and 78 113 relationships. For our experiments, we manually created 8 queries, structurally similar to those in [12], and applied our rewriting technique. The

rewritten queries were evaluated using Cypher.¹ We report our results in Table 1.

4. Conclusion

We presented an algorithm for rewriting NCQs into UC2RPQs over a lightweight ontology that extends DL-Lite with some of the expressive features of \mathcal{ELH} while keeping the data complexity of reasoning in NL. One of our goals is to make progress towards a practicable algorithm, and our prototype implementation, which rewrites the queries into Cypher, suggests that we may be on track. It shows promising results on a real-world dataset from cognitive neuroscience.

Acknowledgements

This work was partially supported by the Federal State of Salzburg under grant number 20102-F2101143-FPR (Digital Neuroscience Initiative) and the Austrian Federal Ministry of Education, Science and Research (BMBWF) under grant number 2920 (Austrian NeuroCloud). This work was also partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

- [1] G. Xiao, L. Ding, B. Cogrel, D. Calvanese, Virtual knowledge graphs: An overview of systems and use cases, *Data Intell.* 1 (2019) 201–223. URL: https://doi.org/10.1162/dint_a_00011. doi:10.1162/DINT_A_00011.
- [2] D. Calvanese, G. D. Giacomo, D. Lembo, M. Lenzerini, R. Rosati, Tractable reasoning and efficient query answering in description logics: The DL-lite family, *Journal of Automated Reasoning* 39 (2007) 385–429. doi:10.1007/s10817-007-9078-x.
- [3] N. Francis, A. Gheerbrant, P. Guagliardo, L. Libkin, V. Marsault, W. Martens, F. Murlak, L. Peterfreund, A. Rogova, D. Vrgoc, A researcher’s digest of GQL (invited talk), in: F. Geerts, B. Vandevoort (Eds.), 26th International Conference on Database Theory, ICDT 2023, March 28–31, 2023, Ioannina, Greece, volume 255 of *LIPICs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023, pp. 1:1–1:22. doi:10.4230/LIPICs.ICDT.2023.1.
- [4] B. Löhnert, N. Augsten, C. Okulmus, M. Ortiz, Towards practicable algorithms for rewriting graph queries beyond dl-lite – full version, 2024. URL: <https://arxiv.org/abs/2405.18181>, accessed: 2024-06-04.
- [5] N. Dragovic, C. Okulmus, M. Ortiz, Rewriting ontology-mediated navigational queries into cypher, in: O. Kutz, C. Lutz, A. Ozaki (Eds.), Proceedings of the 36th International Workshop on Description Logics (DL 2023) co-located with the 20th International Conference on Principles of Knowledge Representation and Reasoning and the 21st International Workshop on Non-Monotonic Reasoning (KR 2023 and NMR 2023), Rhodes, Greece, September 2–4, 2023, volume 3515 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: <https://ceur-ws.org/Vol-3515/paper-9.pdf>.

¹The selected queries happen to be insensitive to the *trail* semantics (implemented in Cypher) vs. the *walk* semantics.

- [6] N. Dragovic, Querying Property Graphs with Ontologies, Master’s thesis, TU Wien, 2022.
- [7] T. Eiter, M. Ortiz, M. Simkus, T. Tran, G. Xiao, Query rewriting for Horn-SHIQ plus rules, in: J. Hoffmann, B. Selman (Eds.), Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada, AAAI Press, 2012, pp. 726–733. URL: <http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4931>.
- [8] N. Dragovic, B. Löhnert, Ontology-mediated querying for property graphs, 2022. URL: <https://gitlab.com/ccns/neurocog/neurodataops/anc/software/owl2cypher>, accessed: 2024-05-24.
- [9] B. Löhnert, B. Engler, F. Hutzler, Cognitive task ontology (COGITO), 2024. URL: <https://gitlab.com/ccns/neurocog/neurodataops/anc/classification/cogito/>, accessed: 2024-05-24.
- [10] R. A. Poldrack, A. Kittur, D. J. Kalar, E. Miller, C. Seppa, Y. Gil, D. S. Parker, F. W. Sabb, R. M. Bilder, The cognitive atlas: Toward a knowledge foundation for cognitive neuroscience, *Frontiers Neuroinformatics* 5 (2011) 17. doi:10.3389/fninf.2011.00017.
- [11] K. Robbins, D. Truong, S. Appelhoff, A. Delorme, S. Makeig, Capturing the nature of events and event context using hierarchical event descriptors (hed), *NeuroImage* 245 (2021) 118766.
- [12] A. Ravensschlag, M. Denissen, B. Löhnert, M. Pawlik, N. A. Himmelstoß, F. Hutzler, Effective queries for mega-analysis in cognitive neuroscience, in: G. Fletcher, V. Kantere (Eds.), Proceedings of the Workshops of the EDBT/ICDT 2023 Joint Conference, Ioannina, Greece, March, 28, 2023, volume 3379 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3379/CoMoNoS_2023_id252_Mateusz_Pawlik.pdf.