NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

as a manuscript

Olga Gerasimova

ONTOLOGY-BASED DATA ACCESS WITH COVERING AXIOMS

PhD Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow – 2023

The PhD Dissertation was prepared at National Research University Higher School of Economics (HSE University).

Academic Supervisor: Sergei O. Kuznetsov, Doctor of Science, Professor, School of Data Analysis and Artificial Intelligence, Faculty of Computer Science, HSE University.

Contents

1 Introduction

Nowadays, ontologies are widely used to improve the convenience of information organisation and access to it in fields such as artificial intelligence, software engineering, biomedical informatics, healthcare, enterprise bookmarking, industrial projects, etc. Answering various types of queries mediated by a description logic (DL) ontology has been known as an essential reasoning problem in knowledge representation since the early 1990s. The proliferation of DLs and their applications, the development of the Web Ontology Language OWL[1], and especially the paradigm of Ontology-Based Data Access have made theory and practice of answering ontology-mediated queries (a pair of ontology and query, for short, OMQs) a hot research area lying at the crossroads of Knowledge Representation and Reasoning, Semantic Technologies and the Semantic Web, Knowledge Graphs, and Database Theory and Technologies.

This thesis aims to explore the practical potential of specific ontology in ontology-mediated query answering tasks and focuses on challenges associated with chosen ontology type. Our study will provide insights into the complication of selected ontology usage for ontology-based data access and analyse opportunities to expand existing practice in this area.

1.1 Object of research

We focus on *Ontology-Based Data Access* (OBDA) [23, 26] with expressive ontologies, where an ontology is used as a helpful tool for supporting query answering for distributed and heterogeneous data sources.

A standard OBDA scenario follows a certain sequence of actions (see Figure 1), where:

- end-user is not involved in the original data organisation;
- user is given an ontology, developed by a domain expert, that defines concepts and properties familiar to the user and provides a vocabulary for the user's queries;

[1] https://www.w3.org/TR/owl2-overview/

– the data schemas (e.g. relational tables) are transformed into data storage in terms of the ontology vocabulary via mapping (encoded in languages such as R2RML – Relational databases (RDB) to Resource Description Framework (RDF) Mapping Language);

– the task of an OBDA system is to convert or *rewrite* user's query, with the help of the ontology and mappings, into an equivalent standard query over the data and answer a new query over the database.

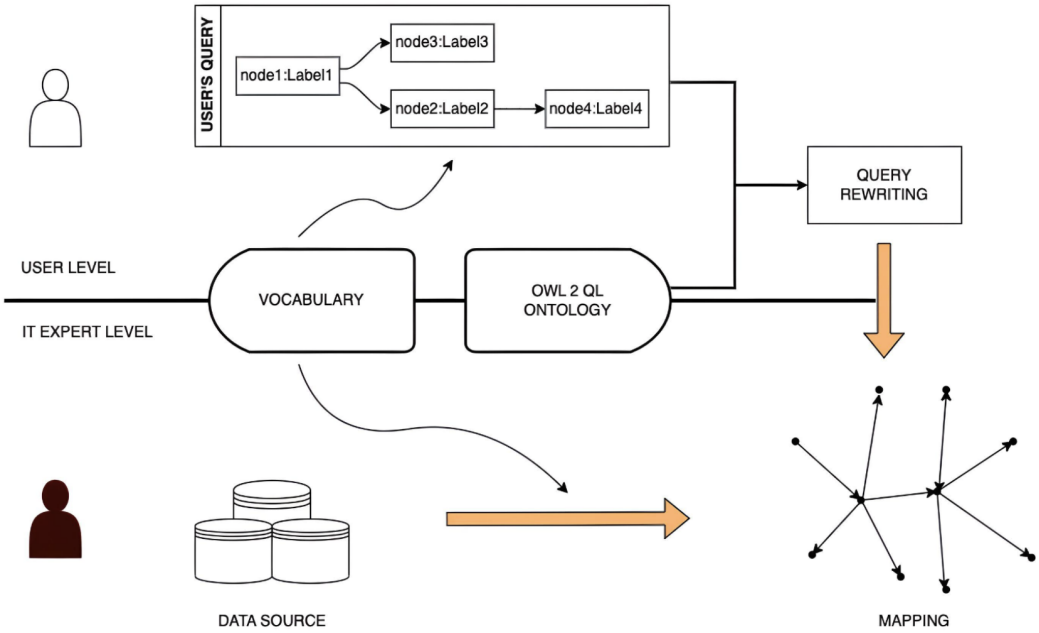


Figure 1: A scheme of OBDA system. The user (domain expert) writes a query with the knowledge of vocabulary and ontology rules describing data dependencies in terms of vocabulary. IT expert transforms the heterogenous data into mappings using given vocabulary. These mappings present data in the graph form consistent with the ontology syntax. Finally, the OBDA system generates a query rewriting into an equivalent standard query over the mappings only.

Thus, an ontology provides a high-level conceptual view of the data, complements the data with background knowledge, and maintains queries over multiple and heterogeneous data sets. For this approach to work, one has to choose the ontology language carefully. In what follows, we mainly consider ontologies formulated regarding suitable description logic.

To ensure theoretical and practical tractability, the OBDA paradigm presupposes that the users' OMQs are reformulated – or rewritten – by the OBDA system into conventional database queries over the original data sources, which have proved to be quite efficiently evaluated by the existing database management systems. Whether or not such a rewriting is possible and into which target query language, naturally depends on the OMQ. One way to *uniformly* guarantee the desired rewritability is to delimit the language for OMQ ontologies and queries. Thus, the *OWL 2 QL* profile[2] of *OWL 2* was designed to guarantee rewritability of *all* OMQs with a *OWL 2 QL* ontology and a conjunctive query (CQ) into first-order (FO) queries, that is, essentially SQL queries. In complexity-theoretic terms, the FO-rewritability of an OMQ means that it can be answered in LogTime uniform AC^0, one of the smallest complexity classes.

In our research, we focus on the problem of determining the complexity and rewritability working for a particular ontology that extends the expressiveness of OBDA. We consider the ontology containing one rule called *covering axiom* – $\{A \sqsubseteq F \sqcup T\}$ – which is not described in the *OWL 2 QL* language. The research based on such an ontology is driven by practical needs since the ontology idea is widely used to describe real data in social networks, industrial processes, decision-making, etc.

1.2 Subject of research

The subject of our research is the covering axiom stating that the union of two other classes covers one class. The relevance of the subject at hand lies in the significant impact on expanding the expressiveness of OBDA in the case that it is unsuitable for standard SQL rewritings over data. The problem of answering ontology-mediated queries with a covering axiom strongly connects graph theory, constraint satisfaction problems, and the other fields of theoretical computer science. The main problem is determining the complexity efficiently and, if it lies in the tractable valuable case for practical application, providing the algorithm for rewriting OMQ into FO query or datalog program. However, even for such a simple small case of one rule in the disjunctive ontology, the computational prob-

[2] https://www.w3.org/TR/owl2-profiles/

lem of answering OMQ remains very complex. We study only Boolean conjunctive queries, because they provide an answer without direct individual item answers due to a covering axiom ambiguity. The tree-shaped form of the queries was chosen in the process of analysing the difficulty of the task at hand and chosen methodology for determining data complexity.

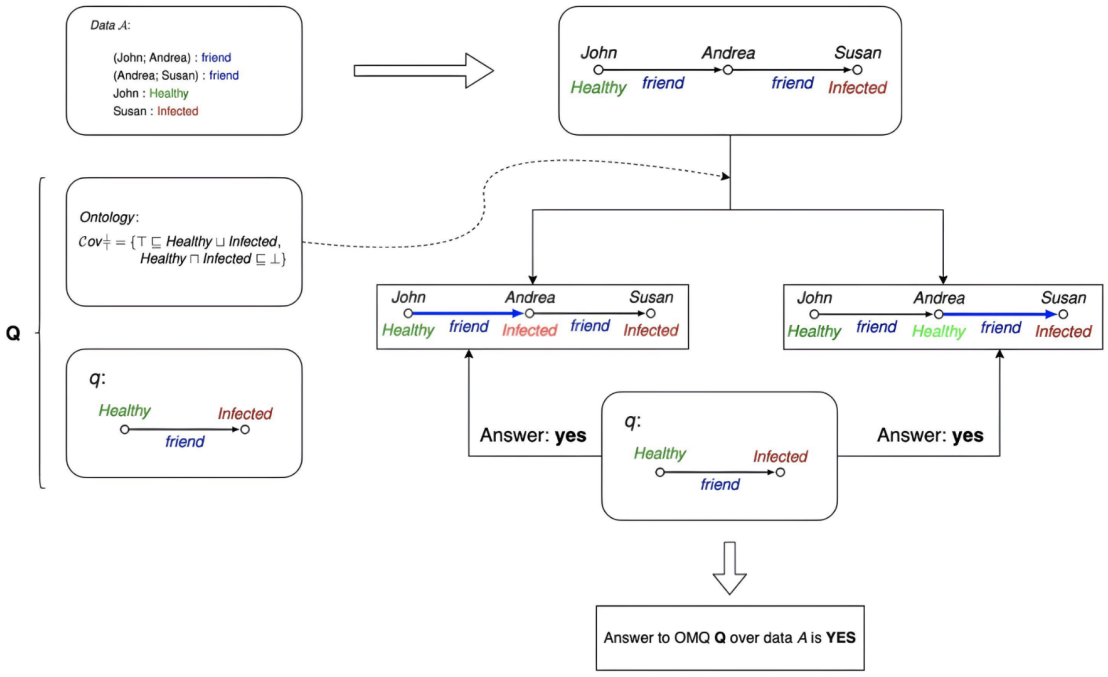


Figure 2: Example of ontology-mediated query Q with a covering axiom. A simple query q has no answers over original data $\mathcal{A}$. However, taking into account all possible label assignments with respect to the ontology, the answer is 'yes' in all the models of the data. Thus, the answer to OMQ is also 'yes'.

To sum up, we focused on Boolean tree-shaped conjunctive queries mediated by a covering axiom. We investigated for them the data complexity and rewritability problems in the framework of OBDA extension. Figure 2 shows an example of ontological query answering.

1.3 Tasks and objectives of the research

The research aims to efficiently determine the data complexity of answering queries mediated by description logic ontologies and construct their optimal rewritings to

standard database queries. The problem is known to be computationally very complex in general without explicit syntactic characterisations available while being extremely important and relevant to ontology-based data access and datalog optimisation. Our study focuses on Boolean conjunctive queries mediated by a simple covering axiom stating that the union of two other classes covers one class.

1.3.1 Data complexity

We study the *data complexity* of answering ontology-mediated Boolean conjunctive queries with a covering axiom:

(d-sirup) $\boldsymbol{Q} = (\mathsf{cov}_A, \boldsymbol{q})$, where $\mathsf{cov}_A = \{ A \sqsubseteq F \sqcup T \}$ is an ontology and $\boldsymbol{q}$ is a Boolean CQ with unary predicates F, T and arbitrary binary predicates.

By the *data complexity*, we determine the computational complexity of answering $\boldsymbol{Q}$ with a specified ontology and a fixed conjunctive query, but over any input data instance $\mathcal{A}$ under the open world semantics.

The main flexible factor that impacts the data complexity in our task is the structure of queries. Thus, we aim to comprehend how the interplay between the *covering axiom* $A \sqsubseteq F \sqcup T$ and the structure of $\boldsymbol{q}$ determines the complexity of $\boldsymbol{Q}$.

Preferring plain graph-theoretic terms for $\boldsymbol{q}$ and data instances because of the convenience of their analysis, manipulation and visualisation as labelled directed graphs, we can formulate the problem of answering $\boldsymbol{Q}$ in the following way:

INSTANCE: any labelled directed graph (digraph, for short) $\mathcal{A}$;

PROBLEM: decide whether each digraph obtained by labelling every A-node in $\mathcal{A}$ with either F or T contains a homomorphic image of $\boldsymbol{q}$ (in which case the certain answer to $\boldsymbol{Q}$ over $\mathcal{A}$ is 'yes').

Formulated above task can be done in coNP [1] as $\boldsymbol{q}$ is *fixed*. So the existence of a homomorphism from $\boldsymbol{q}$ to any labelling of $\mathcal{A}$ can be checked in polynomial time by inspecting all possible $|\mathcal{A}|^{|\boldsymbol{q}|}$-many maps from $\boldsymbol{q}$ to $\mathcal{A}$. To address the given problem, we may consider applying a resolution-based prover or evaluating the disjunctive datalog program $\{(1), (2)\}$ below over $\mathcal{A}$. However, both methods would entail identifying proofs of exponential size.

So, the data complexity problem as the subject of our research leads to the questions:

- whether there exists an alternative, more effective algorithmic solution for a given $\boldsymbol{Q}$ in principle and

- whether it can be executed as a standard (linear, symmetric) datalog or first-order query evaluated over the input graphs $\mathcal{A}$.

1.3.2 Rewritability

Now, it is necessary to raise the issue of the second problem – rewritability – about what we have already mentioned.

By *rewritability*, we assume the reduction of the task of finding certain answers to $\boldsymbol{Q}$ over any input $\mathcal{A}$ to the task of evaluating a conventional database query $\boldsymbol{Q}'$ with optimal data complexity directly over $\mathcal{A}$. The query $\boldsymbol{Q}'$ is then called a *rewriting* of the ontology-mediated query $\boldsymbol{Q}$.

In terms of datalog notations, the OMQ $\boldsymbol{Q} = (\mathsf{cov}_A, \boldsymbol{q})$ is equivalent to the *monadic disjunctive datalog query*

$$T(x) \vee F(x) \leftarrow A(x) \tag{1}$$

$$\boldsymbol{G} \leftarrow \boldsymbol{q} \tag{2}$$

with a nullary (goal) predicate $\boldsymbol{G}$. In the 1980s, trying to understand boundedness (FO-rewritability) and linearisability (linear-datalog-rewritability) of datalog queries, the database community introduced the notion of *sirup* – standing for '*datalog query with a single recursive rule*' [27, 17] – which was thought to be crucial for understanding datalog recursion and optimising datalog programs [21]. Our OMQs $\boldsymbol{Q}$ or disjunctive datalog queries $(\{(1), (2)\}, \boldsymbol{G})$ – which henceforth are referred to as (*monadic*) *disjunctive sirups* (disjunctive sirups) or simply *d-sirups* – play the same fundamental role for understanding OMQs with expressive ontologies and monadic disjunctive datalog queries.

d-sirups may appear syntactically simple, but they actually belong to a highly complex class of OMQs. For example, deciding first-order rewritability of d-sirups turns out to be 2ExpTime-hard – as complicated as deciding program boundedness of arbitrary monadic datalog programs [7, 3]. Interestingly, one of the sources

of this unexpectedly high complexity is the 'twin' FT-labels of nodes in CQs, when atoms $F(x),\ T(x) \in \boldsymbol{q}$, for some variable x. We can eliminate this source by imposing the standard *disjointness constraint* $F \sqcap T \sqsubseteq \bot$ (or $\bot \leftarrow F(x), T(x)$ in datalog parlance), often used in ontologies and conceptual modelling. Thus, we arrive to *dd-sirups* (disjoint disjunctive sirups) of the form

(dd-sirup) $\boldsymbol{Q} = (\mathsf{cov}^{\bot}_{A}, \boldsymbol{q})$, where $\mathsf{cov}^{\bot}_{A} = \{\, A \sqsubseteq F \sqcup T,\ F \sqcap T \sqsubseteq \bot \,\}$.

To make a conclusion of these two subsections, we want to highlight that the complexity and rewritability of both d- and dd-sirups only depend on the structure of the CQs $\boldsymbol{q}$, which suggests a research programme of classifying (d)d-sirups by the type of the graph underlying $\boldsymbol{q}$ – directed path, tree, their undirected variants, etc. – and characterising the data complexity and rewritability of OMQs in the resulting classes.

1.3.3 Comparison of two approaches to query answering via logical reasoners using datalog programs and via data labelling using machine learning approaches

In addition to the theoretical studies, we provide experiments with the help of machine learning to consider our task in terms of the nodes classification problem on graphs. Using state-of-the-art models of graph neural networks, missed data labels can be reconstructed and then we can analyse query answers for different data sets without using an ontology. Taking actual social graphs with nodes labelled by binary classes, we mask labels (removing a part of the original labels) from 5% to 95% labels' coverage and assign labels according to the trained graph neural network. Then, we compare query answers for a graph containing all known labels with the results obtained from our OBDA approaches. Answering Boolean CQs with covering axiom may vary not only from data complexity, which we discussed early, but also from the quantity and the positions of unlabelled data in the graph.

For our OBDA methods, we examine systems that support rule-based approaches for working with disjunctive datalog programs and test them for our task. We compare running time for original task formulation, a query plus ontology, and their rewritable cases.

Finally, we compare two different approaches to identify the profitability of logic reasoning via OMQ rewriting and graph neural network data labelling followed by querying obtained full-labelled graph.

Core **research tasks of Ph.D. research** consist of

- proofing conditions for concrete answering problem for OMQ with a covering axiom to belong to a certain complexity class according to the data size;
- identifying syntactical separation criteria based on query structure;
- finding tractable cases of OMQs with FO- or Datalog- rewritability;
- conducting performance analysis of Datalog-rewritable queries on real data using Datalog reasoning systems;
- providing a comparative analysis of logic reasoners for answering OMQs and their alternatives – machine learning models for labelling data and finding an answer by querying labelled data without the ontology.

The general **goal of our work** is to classify OMQs with a covering axiom based on simple and transparent syntactic conditions on the form of the query and determine OMQs' theoretical and empirical data complexities for the answering task.

1.4 Key results and conclusions

This section describes the main contributions achieved by the present work, its novelty, theoretical and practical significance, research methodology and the reliability of the results.

Key aspects/ideas to be defended:

1. provided reduction of detecting the tractability of a path-OMQ with a covering axiom to checking the tractability of a CSP and tested Polyanna framework working for CSP patterns to find tractable OMQs [15];
2. obtained complete syntactic classification of (d)d-sirups $(\mathsf{cov}_A^{\perp}, \boldsymbol{q})$ with a path-shaped CQ $\boldsymbol{q}$ according to their data complexity and rewritability type [13, 12];

3. performed descriptive analysis on the impact of unlabelled data on the performance of disjunctive datalog reasoners; analysed the efficiency of logic reasoners compared to the use of graph neural networks as a machine learning alternative for ontological query answering problem [16].

Scientific novelty. The originality of our work is unveiled in different aspects. First, the prime research motivation is extending OBDA usage to a more expressive ontology. So, our novel contribution is that we identified tractable cases of queries for ontology with only one covering axiom, where even simple, expressive ontology requires complex developments for its analysis. Secondly, a novel algorithm was suggested for separating tractable cases from intractable ones for a restricted family of queries. The proposed methodology is based on a combination of datalog and automata-theoretic techniques. Also, in researching the practical extension of OBDA with covering axiom, new theoretical results were obtained for establishing computational hardness for different cases. We found several necessary or sufficient conditions on belonging to a particular complexity class or satisfying a specific rewritabily type. We also provided theoretical boundaries on the proposed algorithms for answering ontology-mediated queries with a covering axiom.

Theoretical and practical significance. Nowadays, many existing ontologies fail to comply with the restrictions mandated by the standard languages for OBDA. In practice, the non-complying axioms are often disregarded from the ontology in the hope that there will not be too significant deference between answers to the original and the approximate OMQs.

In response to this problem, our research is an attempt to figure out whether there is another fate for a covering axiom, because it is a widely-used rule in ontologies describing real data. Inspired by the findings of Lutz and Sabellek [25] on a semantic characterisation of OMQs with an OWL 2 EL ontology, we made an important finding for our OBDA extension. Then we developed our own novel techniques serving as a source of inspiration for future theoretical research endeavours in the field and proving the theoretical significance of obtained tetrachotomy of complexity classes in terms of the data complexity for answering OMQs task with a covering axiom.

Practical significance not only lies in the ability to use obtained algorithms for answering OMQs, but also in evaluating whether the logic approach for OMQs can be dominated by the widely adopted machine learning approach for data labelling in the case of various amounts of unlabelled data. The obtained results showed the significance of logical reasoners compared to statistical methods used in the current systems without ontology integration, but also their limitations when the amount of labelled data is too small.

The methodology of the research. The theoretical part of the study is based on computational complexity theory, datalog optimisation, graph theory, automata-theoretic techniques, and description logic. The practical part includes the usage of graph neural networks, classic machine learning and statistics, and, also, testing systems supporting disjunctive datalog.

The reliability of the results. It is ensured by the complete proofs of theorems providing the correctness of results. Practical experiments comprise complex and exhaustive calculations taking into account the metrics' confidence intervals and comparing two diverse approaches such as machine learning-based and logic-based.

Funding. The research was supported by the Faculty of Computer Science, HSE University; Russian Science Foundation; HSE University Basic Research Program; Russian Foundation for Basic Research.

1.5 Publications and approbation of the work

1.5.1 Main publications

1. Gerasimova O., Kikot S., Zakharyaschev M. *Checking the Data Complexity of Ontology-Mediated Queries: A Case Study with Non-uniform CSPs and Polyanna*, in: Description Logic, Theory Combination, and All That. Berlin: Springer, 2019. P. 329-351 [15]

2. Gerasimova O., Kikot S., Kurucz A., Podolskii V. V., Zakharyaschev M. *A Data Complexity and Rewritability Tetrachotomy of Ontology-Mediated Queries with a Covering Axiom*, in: Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning. IJCAI Organization: The International Joint Conference on Artificial Intelligence (IJCAI), 2020. P. 403-413. [13]

3. Gerasimova O., Kikot S., Podolskii V. V., Kurucz A., Zakharyaschev M. *A tetrachotomy of ontology-mediated queries with a covering axiom* // Artificial Intelligence. 2022. Vol. 309. Article 103738. [12]

4. Gerasimova O., Severin N., Makarov I. *Comparative Analysis of Logic Reasoning and Graph Neural Networks for Ontology-Mediated Query Answering with a Covering Axiom* // IEEE Access. 2023. P. 1-13. [16]

1.5.2 Other publications

5. Gerasimova O., Podolskii V. V., Kikot S., Zakharyaschev M. *On the Data Complexity of Ontology-Mediated Queries with a Covering Axiom*, in: Proceedings of the 30th International Workshop on Description Logics, Montpellier, France, July 18-21, 2017. Aachen: CEUR Workshop Proceedings, 2017. Ch. 19. P. 1-12. [11]

6. Gerasimova O., Kikot S., Podolskii V. V., Zakharyaschev M. *More on the Data Complexity of Answering Ontology-Mediated Queries with a Covering Axiom*, in: Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web. Berlin: Springer, 2017. P. 143-158. [14]

7. Gerasimova O., Kikot S., Zakharyaschev M. *Towards a Data Complexity Classification of Ontology-Mediated Queries with Covering*, in: Proceedings of the 31st International Workshop on Description Logics, Tempe, Arizona, October 27-29, 2018. Aachen: CEUR Workshop Proceedings, 2018. P. 1-13. [29]

1.5.3 Reports at conferences and seminars

1. International Workshop Logic Matters, Moscow, Russia, 28 December 2021. Topic: *A Tetrachotomy of Ontology-Mediated Queries with a Covering Axiom.*

2. International Workshop Logic Matters, Moscow, Russia, 29 December 2020. Topic: *Checking the Data Complexity of Ontology-Mediated Queries: A Case Study with Non-uniform CSPs and Polyanna.*

3. The 31st International Workshop on Description Logics, Tempe, Arizona, October 27-29, 2018. Topic: *Towards a Data Complexity Classification of Ontology-Mediated Queries with Covering.*

4. The 8th International Conference on Knowledge Engineering and Semantic Web (KESW), Szczecin, Poland, November 8-10, 2017. Topic: *More on the Data Complexity of Answering Ontology-Mediated Queries with a Covering Axiom.*

5. The 30th International Workshop on Description Logics, Montpellier, France, July 18-21, 2017. Topic: *On the Data Complexity of Ontology-Mediated Queries with a Covering Axiom.*

1.5.4 Personal contribution of the author of the thesis

The author's contribution covers new theoretical conjectures, algorithms and methods developed for classifying theoretical complexity, conducting computational experiments on real-world data, and preparing research papers.

In [15], we formulated the problem of connecting CSP with OBDA for the case when an ontology contains a covering axiom based on [5]. The author of the

thesis used this approach to show that detecting the tractability of a path-OMQ with a covering axiom can be reduced to checking the tractability of a CSP. The author used the Polyanna software, which is based on finding polymorphisms, for checking the computational complexity of CSP patterns corresponding concrete OMQs to detect whether answering a given OMQ with a 4-variable path CQ can be done in P or is coNP-hard. Practical experiments with Polyanna help to define a syntactical condition for P/coNP separation of the OMQs.

In [13, 12], we studied the problem of obtaining transparent syntactic separation criteria for d-sirups with disjoint covering classes and a path-shaped Boolean conjunctive query. The author received several sufficient conditions for membership in AC^0, L, NL, P, and coNP. In particular, the author's contribution was focused on reduction algorithms for L/NL complexity classes via reachability problem. In addition, the author identifies query homomorphisms leading to separation criteria of L/NL and P data complexity classes.

In [16], the author stated the problem of comparing logic approach versus popular machine learning models, such as graph neural networks trained for node classification, to provide an analysis of whether OMQ with a covering axiom can be substituted with label data saturation and then querying directly to data. The author of the thesis formulated the goals for the experiment design, performed logic reasoners evaluation, prepared the paper, and supervised the research on the comparison of logic-based and GNN-based reasoning for OMQs with a covering axiom.

The author of the thesis is a corresponding author in [16] (Q1 WoS, Scopus) and [15] (Scopus). In [13, 12] (Q1 WoS, Scopus), the author of the thesis is placed first, while all the authors significantly contribute to the work.

2 Content of the work

This section provides an overview of the thesis. Each subsection presents original research results in the form of a brief summary of the main ideas, developments, and innovations introduced in the chapter.

Volume and structure of the work. The thesis contains an introductory chapter, a concluding chapter and the content of four papers. In total, the thesis has a page count of 132 pages, including the appendix.

2.1 Checking the Data Complexity of Ontology-Mediated Queries: A Case Study with Non-uniform CSPs and Polyanna

The first chapter of the thesis presents our research on reducing our problem to CSP in order to distinguish between OMQs in P and coNP. In addition, it contains our experience of working with the Polyanna[3] [9] software for finding polymorphisms. The paper's primary goal is to consider the task from a different angle using the methodology of CSPs and, among other things, to summarise what we knew about the data complexity of answering the OMQs for our case study at that moment.

For non-Horn ontology languages (allowing disjunctive axioms), a crucial step in understanding data complexity and rewritability was the discovery in [4] of a connection between OMQs and non-uniform constraint satisfaction problems (CSPs) with a fixed template via MMSNP (Monotone Monadic Strict NP) from [8]. It was used to show that deciding FO- and datalog-rewritability of OMQs with an ontology in any DL between $\mathcal{ALC}$ and $\mathcal{SHIU}$ and an atomic query is NExpTime-complete. The Feder-Vardi dichotomy of CSPs [6, 30] implies a P/coNP dichotomy of such OMQs, which is decidable in NExpTime.

We illustrate how OMQs of the form $(\mathsf{cov}^{\perp}_{\top}, \boldsymbol{q})$ with a path CQ $\boldsymbol{q}$ can be reduced to CSPs (Figure 3). In particular, we are interested in non-uniform CSPs. Let $\mathcal{B}$ be a fixed relational structure called a *template* in this setting. Each template $\mathcal{B}$

[3] ttps://www.cs.ox.ac.uk/activities/constraints/software/#Polyanna

gives rise to the decision problem CSP($\mathcal{B}$) which is to decide, given a data instance $\mathcal{A}$, whether there is a homomorphism from $\mathcal{A}$ to $\mathcal{B}$, in which case we write $\mathcal{A} \to \mathcal{B}$. We show, following [4], how given an OMQ $\boldsymbol{Q} = (\mathsf{cov}^{\perp}_{\top}, \boldsymbol{q})$ with a path CQ $\boldsymbol{q}$, one can construct a template $\mathcal{B}_{\boldsymbol{q}}$ such that, for any data instance $\mathcal{A}$, we have $\mathcal{A} \to \mathcal{B}_{\boldsymbol{q}}$ iff $\mathsf{cov}^{\perp}_{\top}, \mathcal{A} \not\models \boldsymbol{q}$.

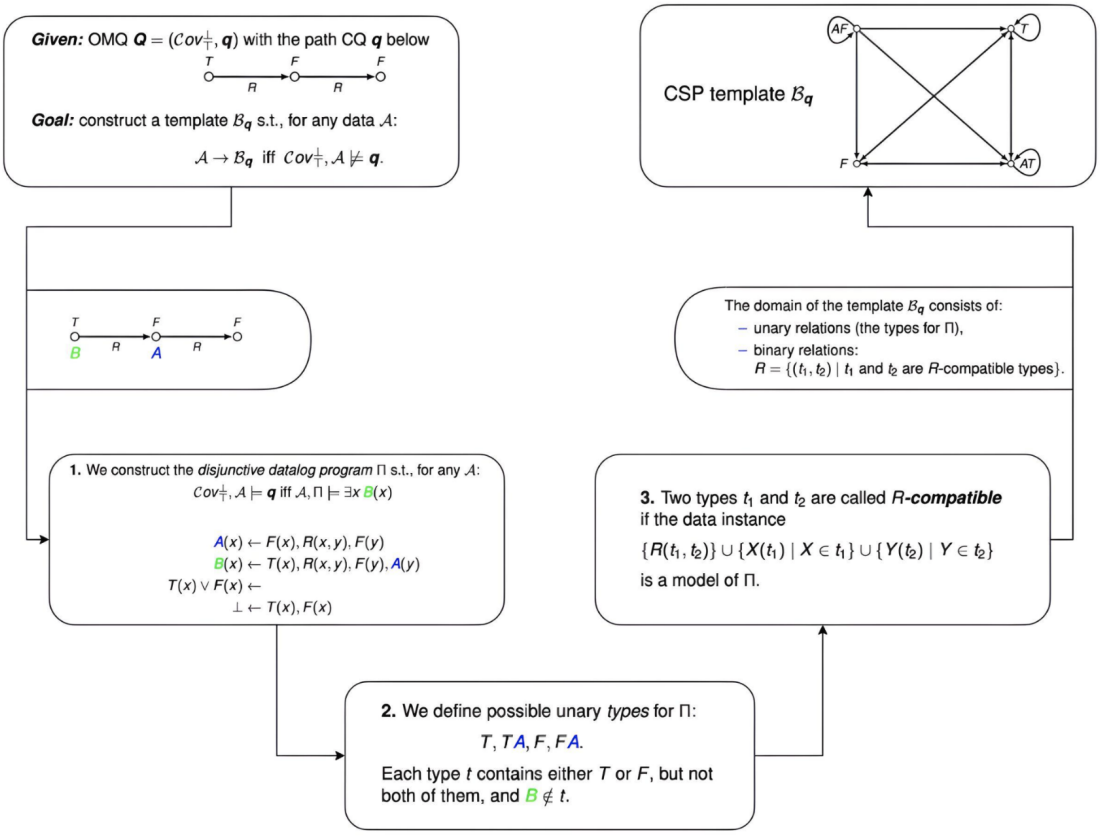


Figure 3: An example of converting OMQs to non-uniform CSPs

To check whether CSP($\mathcal{B}_{\boldsymbol{q}}$) is in P or coNP-hard, one can use the program Polyanna. Polyanna proceeds in two stages. First, it finds a core of the template $\mathcal{B}_{\boldsymbol{q}}$ ('squashing'). Then Polyanna decides tractability or coNP-hardness of CSP($\mathcal{B}_{\boldsymbol{q}}$) by checking whether the core template has polymorphisms of certain types by constructing and solving the corresponding 'indicator problems' [19]. While doing this, Polyanna uses different decomposition techniques to reduce computation in the case when the indicator problem has symmetries. The indicator problem for polymorphisms of arity k and cores with d vertices, for a signature Γ, has $k \cdot d^k$

variables and $\Sigma_{R\in\Gamma}|R|^k$ constraints. Given a core template of size d, Polyanna considers polymorphisms of arity up to $\max(3, d)$. In practice, for our use case, this implies that it can handle cores of size up to 4, but runs out of memory for some cores of size ≥ 5.

Despite the idea that we can transfer results on P/coNP dichotomy for CSPs to our task, we still lack simple and transparent, in particular syntactic, conditions guaranteeing this or that data complexity or type of rewritability. Some results in this direction were obtained in [18, 20] and, within this work, we have presented results on proving coNP-hardness for certain queries with specific structure. However, the fact that a transparent classification of monadic sirups according to their data complexity has not been found so far, and the close connection to CSPs indicates that this problem is extremely hard in general.

The main conclusions of the work are the following. Firstly, we explain by means of a simple example how detecting the tractability of a path-OMQ can be reduced to checking the tractability of a CSP. Secondly, we discuss how the program Polyanna, which was designed to check the tractability of CSPs, can be used in the context of our case study for detecting whether answering a given OMQ with a 4-variable path CQ can be done in P or is coNP-hard. Thirdly, as for the data complexity of answering the OMQs in the framework of our case study, we sketch direct proofs of coNP-hardness using a reduction of 3SAT. Finally, we show how Polyanna can be used for constructing monadic datalog rewritings of tractable OMQs using an arc-consistency check. In addition, in the appendix, we summarise what we know about the data complexity of answering the OMQs for our task.

The novelty of the work is twofold. Firstly, we proposed initial investigations on interconnections between (d)d-sirups and CSPs to transfer results from one formalism to another. Secondly, new theoretical results in terms of the data complexity of answering the OMQs with a covering axiom , where we classify the OMQs according to the number of occurrences of solitary F in their CQs (the case of solitary T is symmetric).

The paper was published in the Springer book "Description Logic, Theory Combination, and All That" indexed in Scopus and WoS.

2.2 A Tetrachotomy of Ontology-Mediated Queries with a Covering Axiom

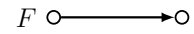
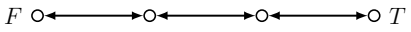
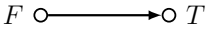
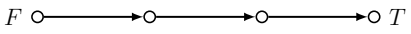
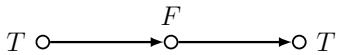
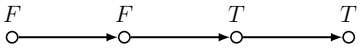
The second chapter contains crucial theoretical results of the research. We obtained some theoretical results for OMQs corresponding to computationally valuable classes from AC^0 to L/NL and P. Queries are described and classified in terms of their structure (node labels, edge labels, edge direction, etc.). The most flexible characteristic is node labelling because ontology works with node labels. Edges' direction and labels can be fixed to simplify the task and clarify the process of node labelling with the help of ontology. Also, the most important result about the complete classification of a restricted family of queries is presented.

In this chapter, we reported on our ongoing attempts to obtain a complete classification of OMQs of the form $\boldsymbol{Q} = (\mathsf{cov}_A, \boldsymbol{q})$, where $\mathsf{cov}_A = \{A \sqsubseteq F \sqcup T\}$, or $\boldsymbol{Q} = (\mathsf{cov}_A^\bot, \boldsymbol{q})$, where $\mathsf{cov}_A^\bot = \{\, A \sqsubseteq F \sqcup T,\ F \sqcap T \sqsubseteq \bot \,\}$, and $\boldsymbol{q}$ is a Boolean CQ. We have observed that answering such OMQs is often tractable, with the respective OMQs being rewritable into standard datalog queries over the data. Sometimes we can even achieve rewritability into linear datalog, which guarantees OMQ answering in NL. We have given a few necessary and sufficient conditions for these phenomena. The simple examples collected in Table 1 show how minor tweaks to $\boldsymbol{q}$ can drastically affect the complexity of OMQs.

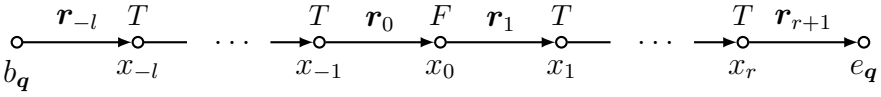
We only consider path CQs $\boldsymbol{q}$ (whose digraph is path-shaped). Solitary F and T nodes will simply be called F- and T-*nodes*, respectively. We denote the first (root) node in $\boldsymbol{q}$ by $b_{\boldsymbol{q}}$ and the last (leaf) node by $e_{\boldsymbol{q}}$. Given nodes x and y, we write $x \prec y$ to say that there is a directed path from x to y in $\boldsymbol{q}$; as usual, $x \preceq y$ means $x \prec y$ or $x = y$. For $x \preceq y$, the set $[x, y]$ comprises those atoms in $\boldsymbol{q}$ whose variables are in the interval $\{z \mid x \preceq z \preceq y\}$ and $(x, y) = [x, y] \setminus \{T(x), F(x), T(y), F(y)\}$. For $\boldsymbol{i} = (x, y)$, we let $|\boldsymbol{i}|$ be the length of the path from x to y, and $|\boldsymbol{q}| = |(b_{\boldsymbol{q}}, e_{\boldsymbol{q}})|$.

We divide path CQs into three disjoint classes: 0-CQs, 1-CQs, and 2-CQs. By a 0-*CQ*, we mean any CQ that does not contain a solitary F (or T, respectively). A 1-*CQ* have exactly one solitary F (or T, respectively) and at least one solitary T (or F, respectively). If a CQ contains at least two F-nodes and at least two T-nodes, it is called 2-CQ. A *twin* in a CQ $\boldsymbol{q}$ is any pair $F(x), T(x) \in \boldsymbol{q}$. dd-sirups $(\mathsf{cov}_A^\bot, \boldsymbol{q})$ with $\boldsymbol{q}$ containing FT-twins are always FO-rewritable.

Table 1: Examples of $\boldsymbol{q}$ from OMQs with the ontology cov_A with respect to their data complexity classes

Complexity	CQ $\boldsymbol{q}$	Explanation
AC^0		if $\boldsymbol{q}$ has only F but no T, then the F can be ignored
L		checks undirected reachability:  the answer to $\boldsymbol{Q}$ is 'yes'
NL		checks directed reachability:  the answer to $\boldsymbol{Q}$ is 'yes'
P		evaluates monotone circuits
coNP		checks CNF satisfiability

We split twinless 1-CQs further into *periodic* and *aperiodic* ones, only considering 1-CQs with a single F-node and at least one T-node (as the case with a single T-node and at least one F-node is symmetric). Given such a twinless 1-CQ $\boldsymbol{q}$ and natural numbers l, r with $l + r \geq 1$, we write $\boldsymbol{q} = \boldsymbol{q}_{lr}$ to say that $\boldsymbol{q}$ has l-many T-nodes $x_{-l} \prec \cdots \prec x_{-1}$ that $\prec$-precede its only F-node x_0, and r-many T-nodes $x_1 \prec \cdots \prec x_r$ that $\prec$-succeed x_0. For every i with $-l \leq i \leq r + 1$, we define a set $\boldsymbol{r}_i$ of binary atoms by taking $\boldsymbol{r}_i = (x_{i-1}, x_i)$, where $x_{-l-1} = b_{\boldsymbol{q}}$ and $x_{r+1} = e_{\boldsymbol{q}}$. Note that $\boldsymbol{r}_i \neq \emptyset$ for $-l < i < r + 1$, but $\boldsymbol{r}_{-l} = \emptyset$ if $b_{\boldsymbol{q}} = x_{-l}$ and $\boldsymbol{r}_{r+1} = \emptyset$ if $x_r = e_{\boldsymbol{q}}$.

$\boldsymbol{q} = \boldsymbol{q}_{lr}$ 

Each $\boldsymbol{r}_i$ determines a finite sequence $\langle \boldsymbol{r}_i \rangle$ of binary predicate symbols. We call $\boldsymbol{q}$ *right-periodic* if $\boldsymbol{q} = \boldsymbol{q}_{0r}$ and either $r = 1$ or $\langle \boldsymbol{r}_i \rangle = \langle \boldsymbol{r}_1 \rangle$ for all $i = 1, \ldots, r$ and $\langle \boldsymbol{r}_{r+1} \rangle = \langle \boldsymbol{r}_1 \rangle^* \lambda$ for some (possibly empty) prefix λ of $\langle \boldsymbol{r}_1 \rangle$. By taking a mirror image of this definition, we obtain the notion of *left-periodic* 1-CQ, in which case $\boldsymbol{q} = \boldsymbol{q}_{l\,0}$ and either $l = 1$ or $\langle \boldsymbol{r}_{-i} \rangle = \langle \boldsymbol{r}_0 \rangle$ for all $i = 1, \ldots, l - 1$ and $\langle \boldsymbol{r}_{-l} \rangle = \lambda \langle \boldsymbol{r}_0 \rangle^*$ for some (possibly empty) suffix λ of $\langle \boldsymbol{r}_0 \rangle$. A twinless 1-CQ $\boldsymbol{q}$ is called *periodic* if it is either right- or left-periodic, and *aperiodic* otherwise.

Theorem ($\mathbf{AC^0}$ / NL / P / coNP-tetrachotomy of (d-sirups) dd-sirups with a (twinlesspath) path CQ). *Let $\boldsymbol{Q}$ be any d-sirup with a twinless path CQ $\boldsymbol{q}$ or any dd-sirup with a path CQ $\boldsymbol{q}$. Then the following tetrachotomy holds (where the three 'if' can be replaced by 'iff' provided that* $\mathrm{NL} \neq \mathrm{P} \neq \mathrm{coNP}$*):*

($\mathbf{AC^0}$) *$\boldsymbol{Q}$ is FO-rewritable and can be answered in* AC^0 *iff $\boldsymbol{q}$ is a 0-CQ or contains an FT-twin; otherwise,*

(NL) *$\boldsymbol{Q}$ is linear-datalog-rewritable and answering it is* NL*-complete if $\boldsymbol{q}$ is a periodic 1-CQ;*

(P) *$\boldsymbol{Q}$ is datalog-rewritable and answering it is* P*-complete if $\boldsymbol{q}$ is an aperiodic 1-CQ;*

(coNP) *answering $\boldsymbol{Q}$ is* coNP*-complete if $\boldsymbol{q}$ is a 2-CQ.*

The main result of our research: a novel complete syntactic classification of dd-sirups $(\mathsf{cov}_A^{\perp}, \boldsymbol{q})$ with a path-shaped CQ $\boldsymbol{q}$ according to their data complexity and rewritability type. While the AC^0/NL part of this AC^0/NL/P/coNP-tetrachotomy follows from our earlier results [11, 14, 29], proving P- and especially coNP-hardness turns out to be tough and requires the development of novel techniques.

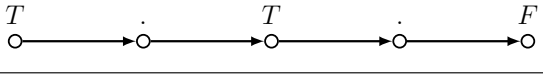
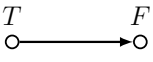
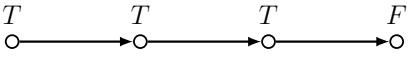
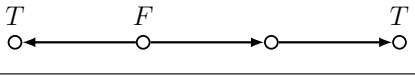
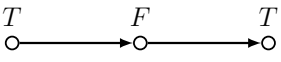
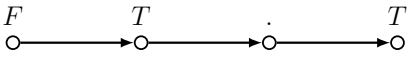
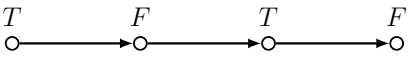
We published two works on these results. The first one was published in the proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (Core A*), and the second one was published in Artificial Intelligence Journal indexed as Q1 (2022) in Scopus and Web of Science.

2.3 Comparative Analysis of Logic Reasoning and Graph Neural Networks for Ontology-Mediated Query Answering with a Covering Axiom

The third chapter focuses on OMQs belonging to the complexity classes L/NL and P, for which there exists a way to rewrite the original query to a datalog program taking into account information from disjunctive ontology.

Table 2 provides the complexity classes and graphical representations of conjunctive queries (adding $\boldsymbol{q}_6$ query from coNP), for which we are conducting experiments on performance evaluation of datalog reasoners in this study.

Table 2: Queries graphical representation as directed labelled graphs

Query	Query graph	Complexity
$\boldsymbol{q}_0$		NL
$\boldsymbol{q}_1$		NL
$\boldsymbol{q}_2$		NL
$\boldsymbol{q}_3$		NL
$\boldsymbol{q}_4$		P
$\boldsymbol{q}_5$		P
$\boldsymbol{q}_6$		coNP

Having produced a datalog program, we aim to understand the efficiency of reasoning solvers for ontology-mediated query answering on different datasets and query patterns. We choose two well-known easy-to-use datalog syntax reasoners DLV (DataLog with $\vee$–disjunction) [24] (via deductive reasoning using disjunctive logic programming), and Clingo [10] (via Answer Set Programming (ASP)). We focused on how the size of unlabelled data impacts the reasoning solvers' performance, which graph neural networks perform the best for node classification, and how querying data enriched with the obtained labelling from GNNs is compared to logic reasoners and ground truth.

We analysed the performance of datalog reasoners depending on the different tractable queries concerning smart/direct datalog rewritings and the percentage of data masked for an ontology evaluation as shown in Figure 4 and Figure 5.

For logic-based reasoners, we have seen that the running time of datalog systems for different sizes of unlabelled data directly depends on the conjunctive query structure. Clingo performs faster than DLV across large networks due to efficient

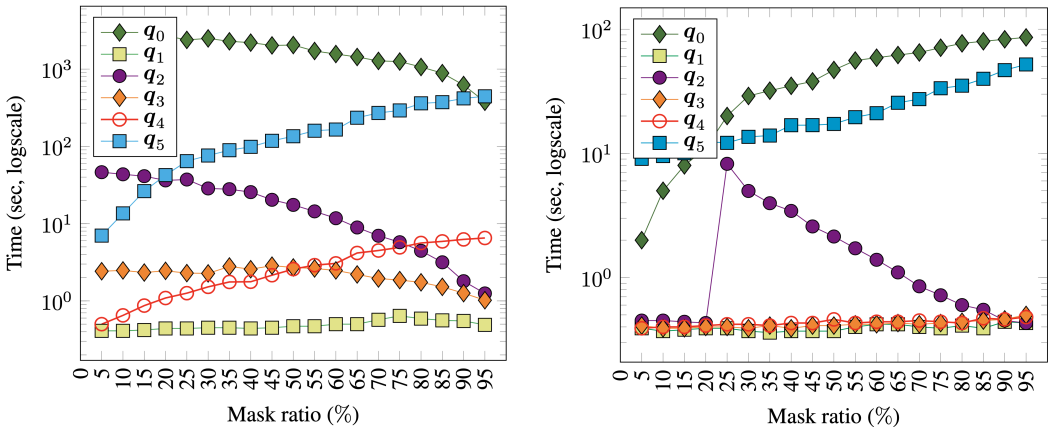


Figure 4: Comparison of the running time of DLV (left) and Clingo (right) systems for different queries on Polblogs.

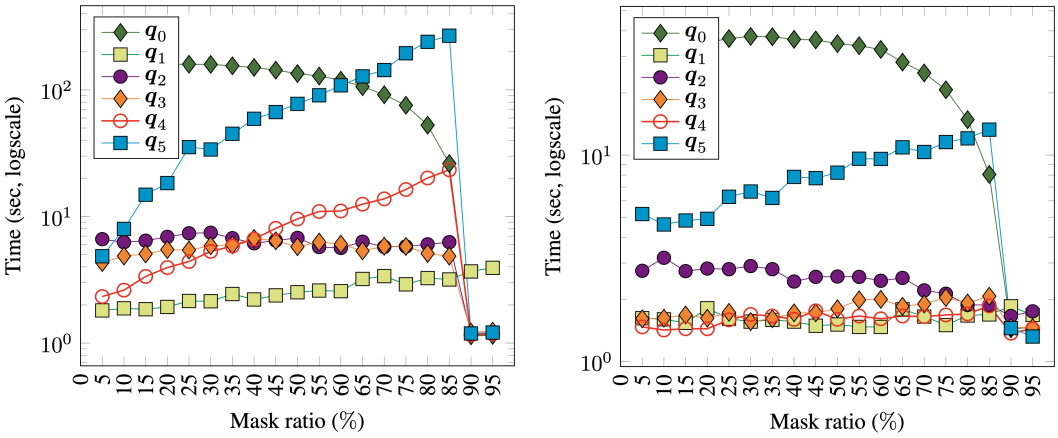


Figure 5: Comparison of the running time of DLV (left) and Clingo (right) systems for different queries on Deezer.

ASP implementation. In addition, it is important to mention that if the size of unlabelled data is too much, then the ontological approach could fail due to a lack of labels, as shown in Figure 5.

For GNN-based reasoning, we have received that the prediction level for node classification, especially for graphs with negative assortativity, is not enough to replace logic reasoners. However, for large networks, even GNN models with high accuracy, such as Graph Convolutional Network(GCN) [22] and Graph Attention Network (GAT) [28], will fail because their node labelling covers most simple OMQs patterns that can be found in the data.

Our results highlight the importance of combined analysis of network and query structures and the amount and placement of unlabelled data for choosing between a precise or approximate decision-making method. The novelty of our work lies in the fact that we have pointed out the limitations of the node classification approach for the OMQ answering problem and have proved the benefits of reasoners and OMQs rewriting in promoting data consistency.

This paper has been published in the IEEE Access Journal. It is rated Q1 (2022) in Web of Science and Scopus.

3 Conclusion

The thesis is based on four published papers: two Q1 journal articles [12, 16], the article in the proceedings of the Core A* conference [13] and the article from the Springer book [15]. The articles [15, 13, 12] provide new approaches in the field of OBDA for a specific case of answering Boolean conjunctive tree-shaped queries mediated by an ontology with a covering axiom. The paper [16] proposes a novel insight into the importance of OBDA compared to the machine learning approach to mitigate ontology rules with data labelling.

All papers together allow us to integrate the covering axiom into OBDA, providing theoretical boundaries of practical OBDA cases and constructive rewriting of OMQs answering into datalog programs. The research bridges the gap between semantic technologies, theoretical computer science and database management.

The main contributions of this thesis to be defended are the following:

1. presented methodology to find tractable cases of a path-OMQ with a covering axiom via checking tractability of a CSP using Polyanna framework and providing monadic datalog rewritings for tractable cases with the complexity AC^0, L/NL, and P.

2. identified complete syntactic classification of (d)d-sirups $(\mathsf{cov}_A^{\perp}, \boldsymbol{q})$ with a path-shaped CQ $\boldsymbol{q}$ according to their data complexity and rewritability type among AC^0 / NL / P / coNP.

3. performed analysis for efficiency of OMQs with a covering axiom comparing OBDA versus machine learning applied for labelling data and directly answering queries over labelled data.

Future research

Directions for future research are raised from restrictions of our results. It is challenging to extend the main theorem on complete classification for more general families of OMQs such as (i) d-sirups with path CQs (that may contain FT-twins), (ii) undirected path-shaped, (iii) ditree- and (iv) undirected tree-shaped dd- and d-sirups. Then, it would be important to settle the tight complexity of deciding FO- and other types of rewritability for arbitrary (d)d-sirups.

The next step is to consider the complexity of deciding a rewritability to more complex ontologies such as Schema.org with multiple disjunctions, *DL-Lite*$_{krom}$ and *DL-Lite*$_{bool}$ [2] with restricted existential quantification on the right-hand side of implications.

Analysing the size of FO-rewritings for OMQs with disjunctive axioms (starting with d- and dd-sirups) is also interesting.

Another direction is considering the data complexity and rewritability problems for (d)d-sirups with multiple answer variables.

In addition, finding lower data complexity bounds for classes of OMQs, for which rewriting algorithms are complete, may be of great importance for closing the case of disjunctive ontologies and their roles in OBDA.

Finally, we need to suggest novel approaches to avoid the effect that saturating data via graph neural networks may provide very precise results for data labelling, but fail to produce correct answers to OMQs.

References

[1] S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[2] A. Artale, D. Calvanese, R. Kontchakov, and M. Zakharyaschev. The DL-Lite family and relations. *Journal of Artificial Intelligence Research (JAIR)*, 36:1–69, 2009.

[3] Michael Benedikt, Balder ten Cate, Thomas Colcombet, and Michael Vanden Boom. The complexity of boundedness for guarded logics. In *30th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2015, Kyoto, Japan, July 6-10, 2015*, pages 293–304. IEEE Computer Society, 2015.

[4] M. Bienvenu, B. ten Cate, C. Lutz, and F. Wolter. Ontology-based data access: A study through disjunctive datalog, CSP, and MMSNP. *ACM Transactions on Database Systems*, 39(4):33:1–44, 2014.

[5] Meghyn Bienvenu, Balder ten Cate, Carsten Lutz, and Frank Wolter. Ontology-based data access: a study through disjunctive datalog, CSP, and MMSNP. In *Proc. of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2013)*, pages 213–224. ACM, 2013.

[6] Andrei A. Bulatov. A dichotomy theorem for nonuniform csps. In Chris Umans, editor, *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 319–330. IEEE Computer Society, 2017.

[7] Stavros S. Cosmadakis, Haim Gaifman, Paris C. Kanellakis, and Moshe Y. Vardi. Decidable optimization problems for database logic programs (preliminary report). In *STOC*, pages 477–490, 1988.

[8] Tomás Feder and Moshe Y. Vardi. The computational structure of monotone monadic SNP and constraint satisfaction: A study through datalog and group theory. *SIAM J. Comput.*, 28(1):57–104, 1998.

[9] Richard Gault and Peter Jeavons. Implementing a test for tractability. *Constraints*, 9:139–160, 2004.

[10] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. Multi-shot asp solving with clingo. *Theory and Practice of Logic Programming*, 19(1):27–82, 2019.

[11] O. Gerasimova, S. Kikot, V. Podolskii, and M. Zakharyaschev. On the data complexity of ontology-mediated queries with a covering axiom. In *Proceedings of the 30th International Workshop on Description Logics*, 2017.

[12] Olga Gerasimova, Stanislav Kikot, Agi Kurucz, Vladimir Podolskii, and Michael Zakharyaschev. A tetrachotomy of ontology-mediated queries with a covering axiom. *Artificial Intelligence*, 309:103738, 2022.

[13] Olga Gerasimova, Stanislav Kikot, Agi Kurucz, Vladimir Podolskii, Michael Zakharyaschev, et al. A data complexity and rewritability tetrachotomy of ontology-mediated queries with a covering axiom. In *Proceedings of the Seventeenth International Conference on Principles of Knowledge Representation and Reasoning*. International Joint Conferences on Artificial Intelligence Organization, 2020.

[14] Olga Gerasimova, Stanislav Kikot, Vladimir V. Podolskii, and Michael Zakharyaschev. More on the data complexity of answering ontology-mediated queries with a covering axiom. In *Knowledge Engineering and Semantic Web - 8th International Conference, KESW 2017, Szczecin, Poland, November 8-10, 2017, Proceedings*, pages 143–158, 2017.

[15] Olga Gerasimova, Stanislav Kikot, and Michael Zakharyaschev. Checking the data complexity of ontology-mediated queries: A case study with non-uniform CSPs and Polyanna. In Carsten Lutz, Uli Sattler, Cesare Tinelli, Anni-Yasmin Turhan, and Frank Wolter, editors, *Description Logic, Theory Combination, and All That*, volume 11560 of *Lecture Notes in Computer Science*, pages 329–351. Springer, 2019.

[16] Olga Gerasimova, Nikita Severin, and Ilya Makarov. Comparative analysis of logic reasoning and graph neural networks for ontology mediated query answering with a covering axiom. *IEEE Access*, pages 1–13, 2023.

[17] Georg Gottlob and Christos H. Papadimitriou. On the complexity of single-rule datalog queries. *Inf. Comput.*, 183(1):104–122, 2003.

[18] André Hernich, Carsten Lutz, Ana Ozaki, and Frank Wolter. Schema.org as a description logic. In Diego Calvanese and Boris Konev, editors, *Proceedings of the 28th International Workshop on Description Logics, Athens,Greece, June 7-10, 2015.*, volume 1350 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[19] Peter Jeavons, David Cohen, and Marc Gyssens. A test for tractability. *CP*, 1118:267–281, 1996.

[20] Mark Kaminski, Yavor Nenov, and Bernardo Cuenca Grau. Datalog rewritability of disjunctive datalog programs and non-Horn ontologies. *Artif. Intell.*, 236:90–118, 2016.

[21] Paris C. Kanellakis. Elements of relational database theory. In Jan van Leeuwen, editor, *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics*, pages 1073–1156. Elsevier and MIT Press, 1990.

[22] N. Kipf, T. and M. Welling. Semi-supervised classification with graph convolutional networks, 2017.

[23] Roman Kontchakov and Michael Zakharyaschev. An introduction to description logics and query rewriting. In *Reasoning Web International Summer School*, pages 195–244. Springer, 2014.

[24] Nicola Leone, Gerald Pfeifer, Wolfgang Faber, Francesco Calimeri, Tina Dell'Armi, Thomas Eiter, Georg Gottlob, Giovambattista Ianni, Giuseppe Ielpa, Christoph Koch, et al. The dlv system. In *Logics in Artificial Intelligence: 8th European Conference, JELIA 2002 Cosenza, Italy, September 23–26, 2002 Proceedings 8*, pages 537–540. Springer, 2002.

[25] Carsten Lutz and Leif Sabellek. Ontology-mediated querying with the description logic EL: trichotomy and linear datalog rewritability. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1181–1187. ijcai.org, 2017.

[26] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Linking data to ontologies. *Journal on Data Semantics*, X:133–173, 2008.

[27] Moshe Y. Vardi. Decidability and undecidability results for boundedness of linear recursive queries. In Chris Edmondson-Yurkanan and Mihalis Yannakakis, editors, *Proceedings of the Seventh ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, March 21-23, 1988, Austin, Texas, USA*, pages 341–351. ACM, 1988.

[28] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks, 2017.

[29] Michael Zakharyaschev, Stanislav Kikot, and Olga Gerasimova. Towards a data complexity classification of ontology-mediated queries with covering. In Magdalena Ortiz and Thomas Schneider, editors, *Proceedings of the 31st International Workshop on Description Logics co-located with 16th International Conference on Principles of Knowledge Representation and Reasoning (KR 2018), Tempe, Arizona, US, October 27th - to - 29th, 2018.*, volume 2211 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

[30] Dmitriy Zhuk. A proof of CSP dichotomy conjecture. In Chris Umans, editor, *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 331–342. IEEE Computer Society, 2017.