On Information Disclosure in Ontology-based Data Access (Extended Abstract)*

Gianluca Cima[1][0000−0003−1783−5605], Domenico Lembo[2][0000−0002−0628−242X], Lorenzo Marconi[2], Riccardo Rosati[2][0000−0002−7697−4958], and Domenico Fabio Savo[3][0000−0002−8391−8049]

[1] University of Bordeaux, CNRS, Bordeaux INP, LaBRI
gianluca.cima@u-bordeaux.fr
[2] Sapienza Università di Roma
{lembo,marconi,rosati}@diag.uniroma1.it
[3] Università degli Studi di Bergamo
domenicofabio.savo@unibg.it

Abstract. This extended abstract summarizes our recent work [4] about Controlled Query Evaluation over Ontology-based data access systems.

Keywords: Ontology-based Data Access · Information Disclosure · Data Protection · First-Order Rewritability

CEUR-WS.org/Vol-2954/abstract-13.pdf

Controlled Query Evaluation (CQE) is an approach to privacy-preserving query answering that recently has gained attention in the context of ontologies [2,6,8,9,12]. In our work, we consider the more general Ontology-based Data Access (OBDA) framework, where an ontology is coupled to external data sources via a declarative mapping [14,15], and extend OBDA with CQE features. In this new framework, which we call *Policy-Protected Ontology-based Data Access (PPOBDA)*, a data protection policy is specified over the ontology of an OBDA specification in terms of logical statements declaring confidential information that must not be revealed to the users. As an example, consider the following formula (expressed as a denial assertion):

$$\forall x, y. OilComp(x) \wedge IssuesLic(x, y) \wedge Comp(y) \rightarrow \bot,$$

which says that the existence of an oil company issuing a license to another company (to operate over its properties) is a private information.

More formally, we define a PPOBDA specification $\mathcal{E}$ as a quadruple $\langle \mathcal{T}, \mathcal{S}, \mathcal{M}, \mathcal{P} \rangle$, where:

- $\mathcal{T}$ is a Description Logic (DL) TBox [1], formalizing intensional domain knowledge;

* Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

- $\mathcal{S}$ is the relational schema at the sources;
- $\mathcal{M}$ is the mapping between $\mathcal{T}$ and $\mathcal{S}$, i.e., a set of logical assertions defining the semantic correspondence between the TBox and the source schema;
- $\mathcal{P}$ is the data protection policy (i.e., a set of formulas) expressed over $\mathcal{T}$.

The components $\mathcal{T}$, $\mathcal{S}$, and $\mathcal{M}$ are exactly as in OBDA specifications, and, as in standard OBDA, a user can only ask queries over the TBox $\mathcal{T}$. Then, query answering is filtered through a *censor*, i.e., a function that alters the answers to queries, in such a way that no data are returned that may lead a malicious user to infer knowledge declared confidential by the policy, even in case she/he accumulates the answers she/he gets over time. Among all possible censors, *optimal* ones are preferred, i.e., those altering query answers in a minimal way.

Within this framework, we initially consider two different notions of censor, called censor in **CQ** and censor in **GA**, previously defined for CQE over DL ontologies [9,12], and which can be naturally extended to PPOBDA. More precisely, given a PPOBDA specification $\mathcal{E} = \langle \mathcal{T}, \mathcal{S}, \mathcal{M}, \mathcal{P} \rangle$, an optimal censor in **CQ** (resp., **GA**) for $\mathcal{E}$ is a function that, taken as input a database instance D for the source schema $\mathcal{S}$, returns a maximal subset $\mathcal{C}$ of the set of Boolean Conjunctive Queries (resp., Ground Atoms) inferred by $\langle \mathcal{T}, \mathcal{S}, \mathcal{M} \rangle$ and D, such that $\mathcal{C} \cup \mathcal{T}$ does not entail information protected by the policy. Since in general several of these maximal sets (incomparable to each other) exist, for both cases we define *query answering under optimal censors* in PPOBDA as a form of skeptical reasoning over all such sets, in the same spirit of [12,6].

Our basic idea to solve query answering under censors is to transform a PPOBDA specification $\mathcal{E}$ into a classical OBDA specification $\mathcal{J}$ (i.e., without policies), in such a way that, whatever database D instantiates the source schema $\mathcal{S}$, query answering under censors in $\mathcal{E}$ over D is equivalent to standard query answering in $\mathcal{J}$ over D. In this transformation, we require that $\mathcal{J}$ has the same TBox of $\mathcal{E}$, so that this reduction is transparent to the user, and the same source schema as $\mathcal{E}$, since, as typical in OBDA, the data sources to be accessed are autonomous. We aim at a transformation independent from the underlying data, so that it can be computed at design-time. This enables us to use off-the-shelf OBDA engines, like Mastro[4] [10] or Ontop[5] [3].

The problem we study can be thus summarized as follows: Given a PPOBDA specification $\mathcal{E} = \langle \mathcal{T}, \mathcal{M}, \mathcal{S}, \mathcal{P} \rangle$, construct an OBDA specification $\mathcal{J} = \langle \mathcal{T}, \mathcal{S}, \mathcal{M}' \rangle$ such that, for any database D for $\mathcal{S}$, conjunctive query answering under optimal censors in $\mathcal{E}$ over D is equivalent to standard conjunctive query answering in $\mathcal{J}$ over D. We investigate this problem for the relevant case in which the TBox is expressed in *DL-Lite*$_{\mathcal{R}}$, the DL underpinning OWL 2 QL [13], and the policy is a set of denial assertions, i.e., conjunctive queries for which an empty answer is imposed due to confidential reasons (as in our initial example). Our contributions are as follows:

[4] http://obdasystems.com/mastro
[5] https://ontop-vkg.org/

(*i*) We show that the above problem has in general no solution when censors in either **CQ** or **GA** are considered, whatever is the DL adopted for expressing the TBox.
(*ii*) To overcome this issue, we propose a further, semantically well-founded approximated notion of censor, named **IGA** (Intersection **GA**) censor, which intuitively, for a PPOBDA specification $\mathcal{E}$ and any database D, returns the intersection of the sets of ground atoms computed by the optimal censors in **GA** for $\mathcal{E}$ applied to D.
(*iii*) We provide an algorithm that solves our problem for every *DL-Lite*$_\mathcal{R}$ PPOBDA specifications under **IGA** censors.
(*iv*) We carried out an experimental evaluation of our approach on (the approximation [7] in *DL-Lite*$_\mathcal{R}$ of) the OBDA NPD benchmark [11]. The tests show that the cost of the transformation performed by our tool is negligible, and answering queries in the presence of a policy in our approach does not cause a significant overhead with respect to the case without policy.

We are currently working on enriching our CQE framework to improve its abilities in the enforcement of confidentiality. In particular, we are investigating more expressive forms of policy, which go beyond denial assertions, and the possibility of expressing preferences that affect the way in which secret information is obfuscated, as in [5].

Acknowledgements. This work was partly supported by the ANR AI Chair INTENDED (ANR-19-CHIA-0014), by the EU within the H2020 Programme under the grant agreement 834228 (ERC Advanced Grant WhiteMec) and the grant agreement 825333 (MOSAICrOWN), by Regione Lombardia within the Call Hub Ricerca e Innovazione under the grant agreement 1175328 (WATCHMAN), and by the Italian MUR (Ministero dell'Università e della Ricerca) through the PRIN project HOPE (prot. 2017MMJJRE), and by Sapienza (project CQEinOBDM).

References

1. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2nd edition, 2007.
2. P. A. Bonatti and L. Sauro. A confidentiality model for ontologies. In *Proc. of the 12th Int. Semantic Web Conf. (ISWC)*, volume 8218 of *Lecture Notes in Computer Science*, pages 17–32, 2013.
3. D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, and G. Xiao. Ontop: Answering SPARQL queries over relational databases. *Semantic Web J.*, 8(3):471–487, 2017.
4. G. Cima, D. Lembo, L. Marconi, R. Rosati, and D. F. Savo. Controlled query evaluation in Ontology-Based Data Access. In *Proc. of the 19th Int. Semantic Web Conf. (ISWC)*, volume 12506, pages 128–146, 2020.
5. G. Cima, D. Lembo, L. Marconi, R. Rosati, and D. F. Savo. Controlled query evaluation over prioritized ontologies with expressive data protection policies. In *Proc. of the 20th Int. Semantic Web Conf. (ISWC)*, 2021.

6. G. Cima, D. Lembo, R. Rosati, and D. F. Savo. Controlled query evaluation in description logics through instance indistinguishability. In *Proc. of the 29th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 1791–1797, 2020.
7. M. Console, J. Mora, R. Rosati, V. Santarelli, and D. F. Savo. Effective computation of maximal sound approximations of description logic ontologies. In *Proc. of the 13th Int. Semantic Web Conf. (ISWC)*, pages 164–179, 2014.
8. B. Cuenca Grau, E. Kharlamov, E. V. Kostylev, and D. Zheleznyakov. Controlled query evaluation over OWL 2 RL ontologies. In *Proc. of the 12th Int. Semantic Web Conf. (ISWC)*, pages 49–65, 2013.
9. B. Cuenca Grau, E. Kharlamov, E. V. Kostylev, and D. Zheleznyakov. Controlled query evaluation for datalog and OWL 2 profile ontologies. In *Proc. of the 24th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 2883–2889, 2015.
10. G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, R. Rosati, M. Ruzzi, and D. F. Savo. MASTRO: A reasoner for effective Ontology-Based Data Access. In *Proc. of the 1st Int. Workshop on OWL Reasoner Evaluation (ORE)*, 2012.
11. D. Lanti, M. Rezk, G. Xiao, and D. Calvanese. The NPD benchmark: Reality check for OBDA systems. In *Proc. of the 18th Int. Conf. on Extending Database Technology (EDBT)*, pages 617–628, 2015.
12. D. Lembo, R. Rosati, and D. F. Savo. Revisiting controlled query evaluation in description logics. In *Proc. of the 28th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 1786–1792, 2019.
13. B. Motik, B. Cuenca Grau, I. Horrocks, Z. Wu, A. Fokoue, and C. Lutz. OWL 2 Web Ontology Language profiles (second edition). W3C Recommendation, World Wide Web Consortium, Dec. 2012. Available at http://www.w3.org/TR/owl2-profiles/.
14. A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Linking data to ontologies. *J. on Data Semantics*, X:133–173, 2008.
15. G. Xiao, D. Calvanese, R. Kontchakov, D. Lembo, A. Poggi, R. Rosati, and M. Zakharyaschev. Ontology-based data access: A survey. In *Proc. of the 27th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 5511–5519, 2018.