Exploiting Ontologies for Explaining Data Sources Semantics

Gianluca Cima[1], Maurizio Lenzerini[1], Antonella Poggi[1,2]

Sapienza Università di Roma
[1]Dipartimento di Ingegneria Informatica, Automatica e Gestionale "Antonio Ruberto"
[2]Dipartimento di Lettere e Culture Moderne
cima, lenzerini, poggi@diag.uniroma1.it

Abstract. We study the problem of associating formal semantic descriptions to data services. We base our proposal on the Ontology-Based Data Access paradigm, where a domain ontology is used to provide a semantic layer mapped to the data sources of an organization. The basic idea is to explain the semantics of a data service in terms of a query over the ontology. We illustrate a formal framework for this problem, based on the notion of source-to-ontology rewriting, which comes in three variants, called sound, complete and perfect, respectively. We present a thorough complexity analysis of two computational problems, namely verification (checking whether a query is a rewriting of a given data service), and computation (computing a rewriting of a data service).

The architecture of many modern Information Systems is based on data services [11], i.e., services deployed on top of data stores, other services, and/or applications to encapsulate a wide range of data-centric operations. In order to realize the promises of data services, in particular to foster their reuse, it is of vital importance to well document and clearly specify their semantics. While most current techniques manually associate APIs (Application Programming Interface) to data services, and describe their intended meaning with ad-hoc methods, often using natural language or complex metadata [3], we propose a new approach, whose goal is to automatically associate formal semantic descriptions to data services. We base our proposal on the *Ontology-Based Data Access* (OBDA) paradigm [9]. An OBDA specification consists of an ontology expressed in Description Logic (DL) [1], the schema of the data sources forming the information system, and a mapping between the source schema and the ontology. The ontology is a formal representation of the underlying domain, and the mapping specifies the relationship between the data at the sources and the concepts in the ontology. The semantics of data services can be thus expressed using the elements of the domain ontology, which is assumed to be familiar to the consumer of data services.

But how can we automatically produce a semantic characterization of a data service, having an OBDA specification available? The idea is to exploit a new reasoning task over the OBDA specification, that works as follows: we express the data service in terms of a query over the sources, and we aim at automatically deriving the query over the ontology that best describes the data service, given the mapping. Note that most of (if not all) the literature about managing data sources through an ontology [7,10] deals with user queries expressed over the ontology, and studies the problem of finding an

ontology-to-source rewriting, i.e., a query over the source schema that, once executed over the data, provides the answers to the original query. Here, the problem is reversed, because we start with a source query and we aim at deriving a corresponding query over the ontology, called a *source-to-ontology rewriting*.

The notions introduced in this paper are relevant in a plethora of scenarios. For the sake of brevity, we mention only two of them. Following the ideas in [4,5], it can be shown that our notions of source-to-ontology rewriting can be used to provide the semantics of open datasets and open APIs published by organizations, which is a crucial aspect for unchaining all the potentials of open data. In [8], the concept of realization of source queries, corresponding to one of the notions studied here, is used for checking whether the mapping provides the right coverage for expressing the relevant data services at the ontology level.

The contributions provided by this work can be summarized as follows. We propose a formal framework for the problem of semantically characterizing a data service through an ontology. We introduce the notions of *perfect*, *sound*, and *complete* source-to-ontology rewritings, and we define two basic reasoning tasks, namely *verification* and *computation*. The former checks whether a given query is a source-to-ontology rewriting of a data service, whereas the latter computes one such rewriting. We show that, although the ideal notion is the one of perfect source-to-ontology rewriting, there are cases where, with the given mapping, no query over the ontology can precisely characterize the data service at hand. Thus, we introduce *maximally sound* and *minimally complete* source-to-ontology rewritings, which intuitively aim at approximating the perfect rewriting of a data service at best, with the goal of either precision (sound rewriting), or recall (complete rewriting).

We study the verification and the computation problem for complete and sound source-to-ontology rewritings in one of the most popular OBDA setting considered in the literature, namely where the ontology language is *DL-Lite*$_{\mathcal{R}}$ [2], each mapping assertion maps a conjunctive query (CQ) over the source to a CQ over the ontology, and both the data service and the source-to-ontology rewriting are expressed as unions of CQs. For complete source-to-ontology rewritings we present algorithms for verification and computation, and characterize the complexity of both tasks. For the case of sound rewritings, we do the same for verification, and we precisely determine the cases where a maximally sound rewriting is not guaranteed to exist.

This discussion paper describes results recently published in [6]. To the best of our knowledge, the problem studied in this work has been (partially) addressed only in [4,8]. The former provides upper bound complexity results for complete rewritings, and the latter focuses on both *DL-Lite*$_{\mathcal{R}}$ and the $\mathcal{EL}$ family of ontology languages, and studies perfect rewritings only, under a slightly different semantics with respect to the one proposed here.

References

1. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press (2003)

2. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. J. of Automated Reasoning **39**(3), 385–429 (2007)
3. Carey, M.J., Onose, N., Petropoulos, M.: Data services. Comm. of the ACM **55**(6), 86–97 (2012)
4. Cima, G.: Preliminary results on ontology-based open data publishing. In: Proc. of DL 2017. CEUR, `ceur-ws.org`, vol. 1879 (2017)
5. Cima, G., Lenzerini, M., Poggi, A.: Semantic technology for open data publishing. In: Proc. of WIMS 2017. p. 1:1 (2017)
6. Cima, G., Lenzerini, M., Poggi, A.: Semantic characterization of data services through ontologies. In: Proc. of IJCAI 2019 (2019)
7. Lenzerini, M.: Managing data through the lens of an ontology. AI Magazine **39**(2), 65–74 (2018)
8. Lutz, C., Marti, J., Sabellek, L.: Query expressibility and verification in ontology-based data access. In: Proc. of KR 2018. pp. 389–398 (2018)
9. Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R.: Linking data to ontologies. J. on Data Semantics **X**, 133–173 (2008). doi:10.1007/978-3-540-77688-8_5
10. Xiao, G., Calvanese, D., an Domenico Lembo, R.K., Poggi, A., Rosati, R., Zakharyaschev, M.: Ontology-based data access: A survey. In: Proc. of IJCAI 2018. pp. 5511–5519 (2018)
11. Zheng, Z., Zhu, J., Lyu, M.R.: Service-generated big data and big data-as-a-service: An overview. In: Proc. of the 2013 IEEE Int. Conf. on Big Data. pp. 403–410 (2013)