Semantic technology for open data publishing

Gianluca Cima[1], Maurizio Lenzerini[1], Antonella Poggi[2]
[1]Dipartimento di Ingegneria Informatica Automatica e Gestionale Antonio Ruberti
[2]Dipartimento di Scienze Documentarie, Linguistico-Filologiche e Geografiche
Sapienza University of Rome, Italy

CCS CONCEPTS

•Information systems → Information integration; •Theory of computation → Description logics; •Applied computing → *Enterprise ontologies, taxonomies and vocabularies;*

ACM Reference format:
Gianluca Cima[1], Maurizio Lenzerini[1], Antonella Poggi[2]. 2016. Semantic technology for open data publishing. In *Proceedings of , , ,* 1 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

After years of focus on technologies for big data storing and processing, many observers are pointing out that making sense of big data cannot be done without suitable tools for conceptualizing, preparing, and integrating data (see http://www.dbta.com/). Data preparation and integration is considered as one of the old problems in data management, and research in the last years has shown that taking into account the semantics of data is crucial for devising powerful data integration solutions. In this work we focus on a specific paradigm for semantic data integration. Indeed, about a decade ago, a new paradigm for modeling and interacting with a data integration systems, called "Ontology-Based Data Access" (OBDA), was proposed [1–4]. According to such paradigm, the client of the information system is freed from being aware of how data and processes are structured in concrete resources (databases, software programs, services, etc.), and interacts with the system by expressing her queries and goals in terms of a conceptual representation of the domain of interest, called ontology. More precisely, a system realizing the vision of OBDA is constituted by three components:

- The *ontology*, whose goal is to provide a formal, clean and high level representation of the domain of interest, and constitutes the component with which the clients of the system (both humans and software programs) interact.
- The *data source* layer, representing the existing data sources in the information system, which are managed by the processes and services operating on their data.
- The *mapping* between the two layers, which is an explicit representation of the relationship between the data sources and the ontology, and is used to translate the operations on the ontology (e.g., query answering) in terms of concrete actions on the data sources.

Thus, OBDA is an advanced approach to semantic data integration, in which the global schema is given in terms of an ontology, i.e., a formal and conceptual view of the application domain, rather than simply a unified view of the data at the sources.

The goal of this work is to provide an overview of the OBDA paradigm, pointing out both the techniques that are at the basis of the paradigm, and some challenges that are currently investigated. In particular, we concentrate on two aspects, *open data publishing*, and *metamodeling*.

The underlying idea is that the OBDA paradigm can provide a formal basis for a principled approach to publish high-quality, semantically annotated open data. The most basic task in open data is the extraction of the content for the dataset(s) to be published, where by "content" we mean both the extensional information (i.e., facts about the domain of interest) conveyed by the dataset, and the intensional knowledge relevant to document such facts (e.g., concepts that intensionally describe facts). In the current methods for open data publishing the semantics of datasets is not formally expressed in a machine-readable form. Conversely, OBDM opens up the possibility of a new way of publishing data, with the idea of annotating data items with the ontology elements that describe them in terms of the concepts in the domain of interest for the organization. When an OBDA system is available in an organization, an obvious way to proceed to open data publication is as follows: (i) express the dataset to be published in terms of a SPARQL query over the ontology, (ii) compute the certain answers to the query, and (iii) publish the result of the certain answer computation, using the query expression and the ontology as a basis for annotating the dataset with suitable metadata expressing its semantics.

Since in open data publishing annotations are expressed in terms of concepts and relations, and asserting and reasoning about annotations is crucial, the need arises to treat concepts and relations as instances of other classes (called metaclasses), and then to use metaclasses as any other elements both in the ontology and in the queries. It follows that metamodeling and metaquerying are essential elements of a methodology for ontology-based open data publishing.

Acknowledgement. This work was supported by the MODEUS project financed by MIUR (SIR programme, grant n. RBSI14TQHQ).

REFERENCES

[1] Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider (Eds.). 2003. *The Description Logic Handbook: Theory, Implementation and Applications.* Cambridge University Press.
[2] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. 2007. Tractable Reasoning and Efficient Query Answering in Description Logics: The *DL-Lite* Family. *J. of Automated Reasoning* 39, 3 (2007), 385–429.
[3] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Riccardo Rosati, and Guido Vetere. 2004. *DL-Lite*: Practical Reasoning for Rich DLs. In *Proc. of the 17th Int. Workshop on Description Logic (DL) (CEUR Electronic Workshop Proceedings*, http://ceur-ws.org/*)*, Vol. 104.
[4] Antonella Poggi, Domenico Lembo, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. 2008. Linking Data to Ontologies. *J. on Data Semantics* X (2008), 133–173. https://doi.org/10.1007/978-3-540-77688-8_5

DOI: 10.1145/nnnnnnn.nnnnnnn