Source Information Disclosure in Ontology-based Data Integration (Extended Abstract)[⋆]

Michael Benedikt, Bernardo Cuenca Grau, and Egor V. Kostylev

Department of Computer Science, University of Oxford, UK

Ontology-based data integration systems allow users to access data sitting in multiple sources by means of queries over a global schema described by an ontology. User queries are formulated against the vocabulary of the ontology and the relationships between the datasources and the ontology terms are specified declaratively by mappings.

In practice, datasources often contain sensitive information that data owners want to keep inaccessible to users. In the setting of ontology-based data integration, the risks of unauthorized information disclosure quickly become apparent since the information exposed to users depends on a complex combination of schema reconciliation, reasoning over the ontology, and access to data in the sources via the mappings.

In this paper, we formalize and study the privacy requirements that a data integration system should satisfy before it is made available to users for querying, as well as on the computational complexity of checking whether such requirements are fulfilled.

Our logical framework for information disclosure builds on work in the database community. In line of existing approaches, we assume that sensitive information is represented by a query (the *policy*) over the source schema, and also that schema-level information (ontology, mappings, source schemas, and policy specification) is publicly available. In contrast, the actual data is only made available as a result of user queries over the ontology. Disclosure in our framework occurs when users are able to uncover an answer to the policy query by querying the system and exploiting the availability of schema-level information. We consider disclosure for a particular dataset, and also whether a schema admits a dataset on which disclosure occurs.

We provide lower and upper bounds on disclosure analysis, in the process introducing a number of techniques for analyzing logical privacy issues in ontology-based data integration. In our analysis, we consider different ontology, mapping, and policy languages. In all cases, we put special emphasis on the results most relevant to standard OBDA, where the ontology is expressed in DL-Lite$_{\mathcal{R}}$ and the mappings are GAV.

Our results have implications on related work. In particular, they imply new lower bounds for the *instance-based determinacy* problem in databases, which is at the core of *data pricing*—the problem of automatically assigning a fair price to a chunk of data given the price of a given set of views.

[⋆] This extended abstract is based on the the paper “Source Information Disclosure in Ontology-based Data Integration”, published in the proceedings of AAAI-2017.