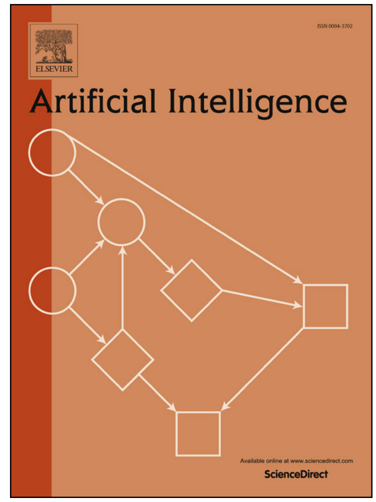
Accepted Manuscript



Logical Foundations of Information Disclosure in Ontology-Based Data Integration

Michael Benedikt, Bernardo Cuenca Grau, Egor V. Kostylev

PII: S0004-3702(18)30306-0
DOI: https://doi.org/10.1016/j.artint.2018.06.002
Reference: ARTINT 3077

To appear in: *Artificial Intelligence*

Received date: 3 July 2017
Revised date: 26 April 2018
Accepted date: 7 June 2018

Please cite this article in press as: M. Benedikt et al., Logical Foundations of Information Disclosure in Ontology-Based Data Integration, *Artif. Intell.* (2018), https://doi.org/10.1016/j.artint.2018.06.002

Logical Foundations of Information Disclosure in Ontology-Based Data Integration

Michael Benedikt, Bernardo Cuenca Grau, Egor V. Kostylev

Department of Computer Science, University of Oxford, UK

Abstract

Ontology-based data integration systems allow users to effectively access data sitting in multiple sources by means of queries over a global schema described by an ontology. In practice, data sources often contain sensitive information that the data owners want to keep inaccessible to users. Our aim in this paper is to lay the logical foundations of information disclosure in ontology-based data integration. Our focus is on the semantic 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. In particular, we formalise and study the problem of determining whether a given data integration system discloses a source query to an attacker. We consider disclosure on a particular dataset, and also whether a schema admits a dataset on which disclosure occurs. We provide matching lower and upper complexity bounds on disclosure analysis, in the process introducing a number of techniques for analysing logical privacy issues in ontology-based data integration.

Keywords: Knowledge Representation and Reasoning, Ontologies, Ontology-based Data Access, Data Integration, Query Answering, Data Privacy.

1. Introduction

Data integration systems expose information from multiple, heterogeneous data sources by means of a *global schema*, in which the mismatches between the individual schemata of the data sources have been reconciled [35, 43, 45, 56, 76]. In addition to reconciling the structure of the data sources, the global schema also enables uniform access to the data by providing users with the vocabulary for query formulation. The relationships between the data sources and the global schema are determined by *mappings*—that is, logical formulae that declaratively specify how each term in the global schema relates to the data in the sources. Queries issued against the global schema are typically answered by one of two approaches. In the first approach, an instance over the global schema is initially materialised using the mappings and the data in the sources; then, the query is answered over the materialised instance. In the second approach, no data is exported from the sources and the global schema remains virtual; this is achieved by first reformulating the user query on-the-fly into a set of queries over the sources, and then assembling back their results.

In *ontology-based data integration* the global schema is extended to an *ontology*—that is, a first-order logic theory expressed in an ontology language such as the Web Ontology Language (OWL) [29]. In addition to defining a high-level conceptual view over the data, the formulae in the ontology also specify how terms in the vocabulary relate to each other, thus providing valuable domain background knowledge that can be exploited to enrich query answers with implicit information.

An important instantiation of the ontology-based data integration paradigm is *ontology-based data access (OBDA)* [23, 68], where the ontology and mapping languages are typically restricted so as to ensure the *first-order rewritability* property: given an arbitrary conjunctive query q over the global schema, ontology $\mathcal{O}$ and mappings $\mathcal{M}$ in the relevant languages, it is possible to reformulate q into a first-order query q' over the source schema such that, for any source data instance $\mathcal{D}$, the answers to q over $\mathcal{O}$, $\mathcal{M}$ and $\mathcal{D}$ coincide with those to q' over $\mathcal{D}$ alone. First-order rewritability therefore ensures that queries can be answered following the second, virtual approach, where the ontology axioms must now also be taken into account during query reformulation. Ontology languages with this property include the QL profile of OWL 2 [65], which is based on the Description Logic DL-Lite$_{\mathcal{R}}$ [22], as well as first-order rule formalisms such as linear tuple-generating dependencies (TGDs) [19, 47]. OBDA has received in recent years a great deal of attention in all its dimensions, including theoretical research [3, 4, 11, 12, 23, 40, 50, 55], system implementation [20, 21, 46], and industrial applications [25, 38, 44, 48, 49].

An important aspect of ontology-based data integration that has so far received only relatively little attention is that of preventing unauthorised information disclosure [24, 26, 31, 66]. In practice, data sources often contain sensitive information, and it is well-known that information integration and linkage poses major threats to the confidentiality of such sensitive data

Email addresses: michael.benedikt@cs.ox.ac.uk (Michael Benedikt), bernardo.cuenca.grau@cs.ox.ac.uk (Bernardo Cuenca Grau), egor.kostylev@cs.ox.ac.uk (Egor V. Kostylev)