Direct Mappings under the Lens of Information Capacity (Extended Abstract)

Davide Lanti[1], Alessandro Mosca[1], Diego Calvanese[1,2] and Marco Montali[1]

[1]Free-University of Bozen-Bolzano, Bolzano, Italy
[2]Umeå University, Umeå, Sweden

Abstract

With the rising popularity of graph-based approaches to data management, exposing the content of traditional, often relational, sources as (knowledge) graphs becomes more and more relevant. In such scenarios, *Direct Mapping* approaches are often used to automatically transform such sources into graph-like formats. A "fundamental" property of these transformations is to be *information preserving*, that is, it should be always possible to (algorithmically) reconstruct the content of the original database. Information preservation, along with other "fundamental" or "desirable" properties proposed in the Semantic Web literature, has never been put into correspondence with over 40 years of extended literature coming from the traditional database perspective. In particular, to the best of our knowledge, it is unknown how classical results on information capacity, dominance, and equivalence, tailored towards specific tasks such as query answering or data update, relate to the results and definitions from the Semantic Web world.

Keywords

Direct Mappings, Information Capacity, Ontology-based Data Access, Virtual Knowledge Graphs

1. Introduction

In the past years, we have been witnessing a renovated interest in graph-like data representation formalisms [1], such as property graphs or RDF graphs, due to the flexibility of their data model when compared to the strict structure of relational databases. Further, graphs from the W3C world are inherently *open-world*, which renders them suitable for publishing data and integrating them with other sources. As these graph structures can often be seen as *OWL ontologies* [2], making use of OWL axioms to formally describe the *knowledge* in a domain of interest, we use the generic term *Knowledge Graphs* (KGs) to refer to all these graph-like data representations formats.

To exploit the advantages of KGs, several companies are publishing their legacy data as graphs, typically by using mapping languages that are inspired by the formalisms studied in the classical literature of Data Integration [3]. A natural question that arose in that community, and that de-facto drives the design of relational databases, is the following: *Given a source schema and a target schema, are they* capable *of representing the same information?* In other words, is it possible to find a bijection between the sets of instances of the two schemas?

DL 2023: 36th International Workshop on Description Logics, September 2–4, 2023, Rhodes, Greece
davide.lanti@unibz.it (D. Lanti); alessandro.mosca@unibz.it (A. Mosca); diego.calvanese@unibz.it (D. Calvanese); marco.montali@unibz.it (M. Montali)
0000-0003-1097-2965 (D. Lanti); 0000-0003-2323-3344 (A. Mosca); 0000-0001-5174-9693 (D. Calvanese); 0000-0002-8021-3430 (M. Montali)

To the best of our knowledge, this question received little attention so far in the Description Logics and Semantic Web scientific literature. There are a few notable exceptions, specifically the work by Arenas et al. [4] or, more recently, the line of works by Thapa et al. [5, 6]. The former work deals with *OWL direct mapping*, that is, with an automated transformation from relational instances to OWL ontologies. The latter work, instead, drops the open-world assumption typical of OWL and considers the closed-world setting of SHACL.

Both lines of works study certain properties of (direct) mappings, like *information preservation* or *semantic preservation*. However, there is no clear connection between these (seemingly) new properties and well-established classical notions on *information capacity* coming from over 40-years of extensive work carried out in the traditional database setting, like [7, 8]. Our aim is to draw such a connection. Furthermore, we are interested in whether notions related to information capacity translate, also for the KG settings under open-world semantics, to common tasks that can be performed over the target schema, such as the ability of querying source data through queries formulated over the target schema, or the ability to perform updates through the target schema [8].

For the reasons above, our aim is to study the problem of mapping relational sources to KGs, trying to re-use traditional notions from the data integration literature. In doing so, one has to address a number of subtleties that arise from an essential mismatch between relational schema and KG semantics: the former are interpreted under the *closed-world assumption* (CWA), whereas the latter are typically interpreted under the *open-world assumption* (OWA). These subtleties are captured by the following questions: *(i)* How to define an appropriate notion of *schema* for KGs, which are typically schema-less? *(ii)* How to define the notion of schema in the presence of mappings but in the absence of an explicit ABox, i.e., for so-called *Virtual Knowledge Graphs* (VKGs)? *(iii)* How do *sound* vs. *exact* mappings affect tasks such as query answering or updates? *(iv)* How does the ontology language of the target KG affect information capacity?

In this work we look at some aspects of the problem, through an example in the virtual setting with OWA, for the update task considering sound vs. exact mappings.

2. Schema Dominance and Taxonomy of Tasks

We now introduce definitions and notions from the information capacity framework of the data integration literature [7, 8]. The general aim of our research is to extend this framework to account for a setting where the target schema is a KG, possibly interpreted under the OWA.

We assume familiarity with relational databases. Given a relational schema Σ, comprehensive of constraints, we denote by $I(\Sigma)$ the set of interpretations (i.e., databases) satisfying Σ.

Definition 1 (Schema Dominance [8]). Given two schemas Σ_1, Σ_2, we say that Σ_2 *dominates* Σ_1, denoted $\Sigma_1 \preceq \Sigma_2$, if there exists a mapping $f : I(\Sigma_1) \to I(\Sigma_2)$ that is total and injective. Schemas Σ_1 and Σ_2 have the *same information capacity*, denoted $\Sigma_1 \equiv \Sigma_2$, if $\Sigma_1 \preceq \Sigma_2$ and $\Sigma_2 \preceq \Sigma_1$. ◁

In principle, arbitrary mappings f may be used to satisfy the above definitions of dominance and equivalence, although non-computable mappings are clearly useless in practice. For this reason, various restrictions have been studied in the database literature [7]. As for the Semantic Web literature, the related notion of *information preservation* [4] (which is not formulated in terms of schema dominance) imposes the (direct) mapping to be a computable function.

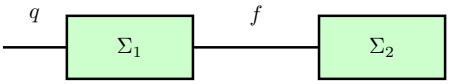

Figure 1: Schemas Σ_1 and Σ_2, with Σ_1 used as an interface for Σ_2.

Information capacity results can be used to determine what kind of operations we can expect to perform on mapped schemas. Assume a situation as in Figure 1, inspired by a tutorial of Atzeni [9]. Miller et al. [8] identified a so-called taxonomy of tasks, depending on how the two schemas are related. If $\Sigma_1 \preceq \Sigma_2$, then one can view the whole DB of Σ_2 through Σ_1. Updates on Σ_2 through Σ_1, instead, can be performed if $\Sigma_2 \equiv \Sigma_1$.

Unfortunately, one can easily show that certain information capacity results cannot be obtained when the target schema is a KG under the OWA, *seemingly* rendering these results not directly applicable to the setting that we are considering here.

3. The Abstract Framework

We briefly discuss some peculiarities of performing updates in a virtual setting, under OWA semantics, and with sound/exact mappings.

A so-called *virtual scenario* is the one typical of *Virtual Knowledge Graphs* (VKGs) [10, 11]. Formally, a *VKG specification* is a triple $\langle \mathcal{T}, \mathcal{M}, \Sigma \rangle$ where $\mathcal{T}$ is an ontology in a lightweight language (e.g., OWL 2 QL), Σ is a relational DB schema, and $\mathcal{M}$ is a set of *mappings* linking the ontology to the data. Figure 2 shows the intuitive semantics of a VKG instance, obtained by pairing the VKG specification with a DB instance $\mathcal{D}$ that satisfies Σ. In VKGs, users issue queries over the *virtual RDF graph* $\langle \mathcal{T}, \mathcal{A}_{\mathcal{M}(\mathcal{D})} \rangle$, realized through the TBox $\mathcal{T}$ and the ABox assertions $\mathcal{A}_{\mathcal{M}(\mathcal{D})}$ obtained by applying $\mathcal{M}$ to $\mathcal{D}$. The ABox is not materialized, but is kept virtual.

In the virtual setting, we consider the target schema to be the VKG specification itself. Hence, we define the set of models $I(\langle \mathcal{T}, \mathcal{M}, \Sigma \rangle)$ of the specification $\langle \mathcal{T}, \mathcal{M}, \Sigma \rangle$, as the set of models of the VKG $\langle \mathcal{T}, \mathcal{A}_{\mathcal{M}(\mathcal{D})} \rangle$, for any database instance $\mathcal{D}$ satisfying Σ. Note that this definition implies that virtual ABoxes in a VKG are not arbitrary, but must be generated by applying the mapping assertions in $\mathcal{M}$ to some valid database instance $\mathcal{D}$ of Σ. This poses an additional constraint with respect to the non-virtual setting (where one would instead allow for arbitrary ABoxes): any update modifying a virtual ABox should still produce a virtual ABox. There are different possibilities for defining a virtual ABox. Interestingly, these do not admit the same updates. A VKG setting uses so-called *global-as-view* (GAV) mappings [3, 12], which are assertions of the form $g \rightsquigarrow q_{\mathcal{S}}$, where g is an atom over a TBox predicate[1] and $q_{\mathcal{S}}$ is a query over the source schema. The form of the virtual ABox $\mathcal{A}_{\mathcal{M}(\mathcal{D})}$ depends on how such mapping assertions in $\mathcal{M}$ are interpreted. Intuitively, if the mapping assertions are *exact* [3], then $\mathcal{A}_{\mathcal{M}(\mathcal{D})}$ must contain *exactly* those atoms derivable through the mappings from the database. Instead, if the mappings are *sound* [3], then $\mathcal{A}_{\mathcal{M}(\mathcal{D})}$ could contain more atoms than those strictly required by the mapping assertions, provided that the semantics of the VKG does not change. Hence, the only meaningful way of adding more atoms is allowing only those that are semantically entailed by $\mathcal{T}$. Such VKGs

[1] The atom may contain so-called "template" functions used to construct identifiers of objects in the KG.

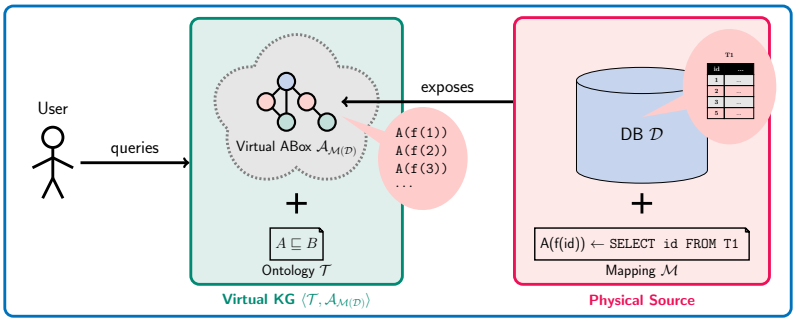

Figure 2: VKG Scenario. The user poses queries over the VKG $(\mathcal{T}, A_{\mathcal{M}(\mathcal{D})})$.

are sometimes called *saturated*, and are semantically equivalent to their "exact" counterpart.

It is easy to see that considering sound or exact mappings, even when not impacting the semantics of the VKG, does still have an impact w.r.t. updates.

Example 2. Consider the scenario in Figure 2, and assume we want to *add* $B(f(1))$ *to the virtual ABox*. Since $B(f(1))$ is already entailed by $\langle \mathcal{T}, \mathcal{A}_{\mathcal{M}(\mathcal{D})} \rangle$, this is a seemingly reasonable update to do (checking entailments a-priori is unrealistic for huge, bulk updates). However, assuming exact mappings, $\mathcal{A}_{\mathcal{M}(\mathcal{D})} \cup \{B(f(1))\}$ is not a valid virtual ABox for the VKG specification $\langle \mathcal{T}, \mathcal{M}, \Sigma \rangle$. Observe that the same update can instead be performed if sound mappings are considered. ◁

It is possible to provide an alternative VKG specification that still uses exact mappings, but that allows for the update in the example above. The trick is to exploit the technique of *saturated mappings* [13, 14], that essentially "compiles" the axioms in the ontology in the mappings $\mathcal{M}$.

Example 3. Consider a modified version of the scenario in Example 2, where the VKG specification contains an additional mapping $B(f(id)) \rightsquigarrow$ `SELECT id FROM T1`, and the ontology is empty. It is easy to observe that the new VKG specification is semantically equivalent to the original one. Further, it is now possible to perform the update, even when mappings are assumed to be exact, since $B(f(1))$ is already present in the virtual ABox. ◁

4. Future Work

Currently, we are looking into the relationship between notions such as information or semantic preservation from Semantic Web literature and their counterparts in terms of schema capacity. Our goal is to provide a full framework able to characterize and compare generic GAV-based Direct Mapping strategies.

Acknowledgments

This research is supported by the Province of Bolzano through the project D2G2, by EURAC Research (Italy) through the project CRISP, and by the Free University of Bozen-Bolzano through the project MP4OBDA. Diego Calvanese is supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutiérrez, S. Kirrane, J. E. Labra Gayo, R. Navigli, S. Neumaier, A.-C. Ngonga Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge Graphs, Synthesis Lectures on Data, Semantics, and Knowledge, Springer, 2022. doi:10.1007/978-3-031-01918-0.

[2] J. Bao, et al., OWL 2 Web Ontology Language Document Overview (Second Edition), W3C Recommendation, World Wide Web Consortium, 2012. Available at http://www.w3.org/TR/owl2-overview/.

[3] M. Lenzerini, Data integration: A theoretical perspective., in: Proc. of the 21st ACM Symp. on Principles of Database Systems (PODS), 2002, pp. 233–246. doi:10.1145/543613.543644.

[4] J. F. Sequeda, M. Arenas, D. P. Miranker, On directly mapping relational databases to RDF and OWL, in: Proc. of the 21st Int. World Wide Web Conf. (WWW), ACM, 2012, pp. 649–658. doi:10.1145/2187836.2187924.

[5] R. B. Thapa, M. Giese, A source-to-target constraint rewriting for direct mapping, in: Proc. of the 20th Int. Semantic Web Conf. (ISWC), volume 12922 of *LNCS*, Springer, 2021, pp. 21–38. doi:10.1007/978-3-030-88361-4_2.

[6] R. B. Thapa, M. Giese, Mapping relational database constraints to SHACL, in: Proc. of the 21st Int. Semantic Web Conf. (ISWC), volume 13489 of *LNCS*, Springer, 2022, pp. 214–230. doi:10.1007/978-3-031-19433-7_13.

[7] R. Hull, Relative information capacity of simple relational database schemata, in: Proc. of the 3rd ACM Symp. on Principles of Database Systems (PODS), ACM, 1984, pp. 97–109. doi:10.1145/588011.588027.

[8] R. J. Miller, Y. E. Ioannidis, R. Ramakrishnan, The use of information capacity in schema integration and translation, in: Proc. of the 19th Int. Conf. on Very Large Data Bases (VLDB), Morgan Kaufmann, 1993, pp. 120–133. URL: http://www.vldb.org/conf/1993/P120.PDF.

[9] P. Atzeni, Schema and data translation, in: Proc. of the 22nd IEEE Int. Conf. on Data Engineering (ICDE), 2006, p. 103. doi:10.1109/ICDE.2006.134.

[10] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati, Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family, J. of Automated Reasoning 39 (2007) 385–429. doi:10.1007/s10817-007-9078-x.

[11] G. Xiao, D. Calvanese, R. Kontchakov, D. Lembo, A. Poggi, R. Rosati, M. Zakharyaschev, Ontology-based data access: A survey, in: Proc. of the 27th Int. Joint Conf. on Artificial Intelligence (IJCAI), IJCAI Org., 2018, pp. 5511–5519. doi:10.24963/ijcai.2018/777.

[12] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati, Linking data to ontologies, J. on Data Semantics 10 (2008) 133–173. doi:10.1007/978-3-540-77688-8_5.

[13] M. Rodriguez-Muro, R. Kontchakov, M. Zakharyaschev, Ontology-based data access: Ontop of databases, in: Proc. of the 12th Int. Semantic Web Conf. (ISWC), volume 8218 of *LNCS*, Springer, 2013, pp. 558–573. doi:10.1007/978-3-642-41335-3_35.

[14] J. F. Sequeda, M. Arenas, D. P. Miranker, OBDA: Query rewriting or materialization? In practice, both!, in: Proc. of the 13th Int. Semantic Web Conf. (ISWC), volume 8796 of *LNCS*, Springer, 2014, pp. 535–551. doi:10.1007/978-3-319-11964-9_34.