Mapping Repair in Ontology-based Data Access Evolving Systems (Extended Abstract)

Domenico Lembo[1], Riccardo Rosati[1], Valerio Santarelli[1],
Domenico Fabio Savo[1], Evgenij Thorstensen[2]

[1] Sapienza Università di Roma
lastname@dis.uniroma1.it

[2] University of Oslo
evgenit@ifi.uio.no

Introduction. *Ontology-based data access (OBDA)* is the problem of accessing source databases through the mediation of a conceptual domain view, given in terms of an ontology [9]. An *OBDA specification* is constituted by an ontology, usually a Description Logic (DL) TBox [2], a schema of the source databases, and a mapping specifying the relationship between the source data and the elements of the ontology, which is commonly given as a set of assertions, each one associating a query over the source schema with a query over the ontology. In the following, we denote an OBDA specification as $\mathcal{J} = \langle \mathcal{T}, \mathcal{S}, \mathcal{M} \rangle$, where $\mathcal{T}$ is the TBox, $\mathcal{S}$ is the source schema, and $\mathcal{M}$ is the mapping.

A major issue in OBDA concerns the design and management of a specification, and in particular of a mapping [6]. Mapping design is indeed a time-consuming and complex operation, which in general cannot be totally automatized. Of course, modifying the mapping due to changes in the other components of the specification may result in a time-consuming process as well. We experienced this in various industrial and academic projects. Among them, we mention a collaboration with the Italian Ministry of Economy and Finances, where we had to map a domain ontology with a source database completely independent from it, which caused mapping definition to be particularly complex [1], and the use cases of the EU project Optique, focused on OBDA for Big Data [4], which were particularly challenging with respect to mapping design and analysis.

In this paper, we study the evolution of OBDA specifications, and focus on the typical case in which the ontology and/or the source schema are updated and the mapping needs to be in turn modified to restore system consistency. Our approach is based on the idea of repairing the mapping according to the usual principle of minimal change and on a recent, mapping-based notion of consistency of the specification. We define and analyze two notions of mapping repair, called deletion-based and entailment-based repair, respectively. Whereas the former notion adopts the simple idea of looking only at consistent subsets of the original mapping, the latter aims at preserving as much as possible of the mapping assertions that are entailed by the initial specification and do not contradict the update. We then present a set of initial results on the complexity of query answering in OBDA under ontology update in both the repair semantics.

We observe that many approaches exist for both *ontology evolution* [11] and *database schema evolution* [10]. However, to the best of our knowledge, no previous study has analyzed evolution in the presence of a mapping connecting an ontology to a database schema. In this sense, a problem that is close to OBDA is *ontology matching and alignment*, which is based on the use of a notion of mapping to integrate different ontologies. Several works have studied the problem of repairing inconsistent mappings in this context (e.g., [5, 8]), but the framework considered is very different from OBDA.

The present paper is an extended abstract of [7].

OBDA specifications. A source schema $\mathcal{S}$ is a relational schema, possibly equipped with integrity constraints (ICs). A legal instance for $\mathcal{S}$ is a database that satisfies the ICs in $\mathcal{S}$. We say that a source schema is *simple* if it has no ICs.

A *mapping assertion* m from a source schema $\mathcal{S}$ to a TBox $\mathcal{T}$ has the form $\phi(\boldsymbol{x}) \rightsquigarrow \psi(\boldsymbol{x})$, where $\phi(\boldsymbol{x})$ and $\psi(\boldsymbol{x})$ are queries over $\mathcal{S}$ and $\mathcal{T}$, respectively, both with free variables $\boldsymbol{x}$. A mapping $\mathcal{M}$ from $\mathcal{S}$ to $\mathcal{T}$ is a *possibly infinite* set of mapping assertions from $\mathcal{S}$ to $\mathcal{T}$. We consider the notable cases in which $\phi(\boldsymbol{x})$ is a CQ, and $\psi(\boldsymbol{x})$ is either a CQ without constants (GLAV mapping language), or a single-atom query without constants and existential variables (GAV mapping language).

An OBDA specification is a triple $\mathcal{J} = \langle \mathcal{T}, \mathcal{S}, \mathcal{M} \rangle$, where $\mathcal{T}$ is a TBox, $\mathcal{S}$ is a source schema, and $\mathcal{M}$ is a mapping between the two. The semantics of $\mathcal{J}$ is given with respect to a legal instance D for $\mathcal{S}$: a model for $\mathcal{J}$ w.r.t. D is a first-order interpretation $\mathcal{I}$ that satisfies $\mathcal{T}$ and *satisfies* $\mathcal{M}$ *w.r.t.* D, i.e., if for each assertion $\phi(\boldsymbol{x}) \rightsquigarrow \psi(\boldsymbol{x})$ in $\mathcal{M}$ and each tuple of constants $\boldsymbol{t}$ such that $\boldsymbol{t}$ is in the evaluation of $\phi(\boldsymbol{x})$ over D, we have that $\mathcal{I} \models \psi(\boldsymbol{t})$. If $\mathcal{J}$ has no models w.r.t. D, we say that $(\mathcal{J}, D)$ is *inconsistent*. Also, we denote denote with $(\mathcal{J}, D) \models \alpha$ the entailment of a sentence α by $(\mathcal{J}, D)$.

DMR and EMR repairs. We adopt a *mapping-centered* notion of OBDA evolution: given an OBDA specification $\mathcal{J} = \langle \mathcal{T}, \mathcal{S}, \mathcal{M} \rangle$, we want to repair the mapping $\mathcal{M}$ given a modification of the TBox $\mathcal{T}$ and/or of the source schema $\mathcal{S}$. This is a natural assumption: indeed, the mapping is an information that depends on both the TBox and the source schema, while the TBox and the schema are (at least in principle) semantically independent entities (since the data sources are autonomous systems).

Following the classical approaches to belief revision, we want to find a notion of repair of a mapping that is based on two general principles: (i) it should preserve *consistency* of the OBDA specification; (ii) it should express *minimal change* with respect to the initial OBDA specification. With respect to consistency preservation, we adopt a non-classical notion of inconsistency for an OBDA specification, called *global mapping inconsistency*, recently introduced in [6]. According to this notion, a mapping $\mathcal{M}$ is globally inconsistent for $\langle \mathcal{T}, \mathcal{S} \rangle$, with $\mathcal{T}$ a TBox and $\mathcal{S}$ a source schema, if there exists no instance D for $\mathcal{S}$ that *activates* all the mapping assertions in $\mathcal{M}$ and such that $(\langle \mathcal{T}, \mathcal{S}, \mathcal{M} \rangle, D)$ is consistent. We say that D activates a mapping assertion $\phi(\boldsymbol{x}) \rightsquigarrow \psi(\boldsymbol{x})$ if $\phi(\boldsymbol{x})$ has a non-empty answer on D. In the context of OBDA, global mapping inconsistency provides a more meaningful notion of inconsistency than the classical one (which considers all possible source instances): for example, in all the cases when $\mathcal{S}$ is a relational schema with standard ICs, the OBDA specification is inconsistent according to the classical semantics iff its TBox is inconsistent.

With respect to minimal change, we propose two different notions of repair. The first one, called *deletion-based mapping repair (DMR)*, reflects the simple idea of repairing a mapping through a (subset-)minimal deletion of assertions from the initial mapping.

Let $\mathcal{J} = \langle \mathcal{T}, \mathcal{S}, \mathcal{M} \rangle$ be an OBDA specification such that $\mathcal{M}$ is globally consistent for $\langle \mathcal{T}, \mathcal{S} \rangle$, $\mathcal{T}'$ a consistent TBox, $\mathcal{S}'$ a consistent source schema, and $\mathcal{M}'$ a mapping such that $\mathcal{M}' \subseteq \mathcal{M}$. We say that $\mathcal{M}'$ is a *DMR for* $\mathcal{J}$ *under update* $\langle \mathcal{T}', \mathcal{S}' \rangle$ if: (i) $\mathcal{M}'$ is globally consistent for $\langle \mathcal{T}', \mathcal{S}' \rangle$; and (ii) there exists no mapping $\mathcal{M}'' \subseteq \mathcal{M}$ such that

$\mathcal{M}''$ is globally consistent for $\langle \mathcal{T}', \mathcal{S}' \rangle$, and $\mathcal{M}'' \supset \mathcal{M}'$. In other words, a DMR is a maximal subset of $\mathcal{M}$ that is globally consistent for the new TBox and source schema.

It is easy to see that a DMR for $\mathcal{J}$ under update $\langle \mathcal{T}', \mathcal{S}' \rangle$ always exists, and that if $\mathcal{M}$ is finite, also $\mathcal{M}'$ is finite. Furthermore, if $\mathcal{M}$ is globally consistent for $\langle \mathcal{T}', \mathcal{S}' \rangle$, then $\mathcal{M}$ is the only DMR for $\mathcal{J}$. In general, however, several repairs exist.

We are now able to introduce our DMR-based update operator, denoted with $\bullet$, and define the OBDA specifications computed by such operator. Given $\mathcal{J}$ under update $\langle \mathcal{T}', \mathcal{S}' \rangle$, the set of such specifications, denoted by $\mathcal{J} \bullet \langle \mathcal{T}', \mathcal{S}' \rangle$, is

$$\{\langle \mathcal{T}', \mathcal{S}', \mathcal{M}' \rangle \mid \mathcal{M}' \text{ is a DMR for } \mathcal{J} \text{ under update } \langle \mathcal{T}', \mathcal{S}' \rangle\}.$$

Note that the above notions are actually independent of the initial TBox $\mathcal{T}$ and schema $\mathcal{S}$. Therefore, in the following we will simply call $\mathcal{M}'$ a DMR for $\mathcal{M}$ under update $\langle \mathcal{T}', \mathcal{S}' \rangle$, and will denote the set $\mathcal{J} \bullet \langle \mathcal{T}', \mathcal{S}' \rangle$ as $\mathcal{M} \bullet \langle \mathcal{T}', \mathcal{S}' \rangle$.

The notion of query entailment in the DMR-based update framework is as follows. Let $\mathcal{M}$ be a mapping, $\mathcal{T}'$ a consistent TBox, $\mathcal{S}'$ a consistent source schema, D a legal instance for $\mathcal{S}'$, and q a Boolean conjunctive query (BCQ). We say that *q is entailed under DMR by $\mathcal{M}$, $\mathcal{T}'$, $\mathcal{S}'$, and D*, denoted as $(\mathcal{M} \bullet \langle \mathcal{T}', \mathcal{S}' \rangle, D) \models q$, if $(\mathcal{J}', D) \models q$ for every $\mathcal{J}' \in \mathcal{M} \bullet \langle \mathcal{T}', \mathcal{S}' \rangle$.

The second notion of repair, called *entailment-based mapping repair (EMR)*, relies on the *mapping entailment set (MES)*. The MES of an OBDA specification $\mathcal{J}$ for a mapping language $\mathcal{L}$, denoted $MES_{\mathcal{L}}(\mathcal{J})$, is a set of mapping assertions in $\mathcal{L}$ that are logical consequences of $\mathcal{J}$, and such that their body coincides with the body of an assertion in the original mapping. Then, the repairs are globally consistent mappings that allow for preserving as much as possible of the initial MES, according to a minimality criterion that formalizes the intuitive principle of preferring insertions over deletions. Formally, $\mathcal{M}'$ is an *EMR for $\mathcal{J}$ under update* $\langle \mathcal{T}', \mathcal{S}' \rangle$ if: (i) $\mathcal{M}'$ is globally consistent for $\langle \mathcal{T}', \mathcal{S}' \rangle$; and (ii) there exists no $\mathcal{L}$-mapping $\mathcal{M}''$ such that $\mathcal{M}''$ is globally consistent for $\langle \mathcal{T}', \mathcal{S}' \rangle$ and $MES_{\mathcal{L}}(\langle \mathcal{T}', \mathcal{S}', \mathcal{M}'' \rangle)$ has *fewer changes* than $MES_{\mathcal{L}}(\langle \mathcal{T}', \mathcal{S}', \mathcal{M}' \rangle)$ with respect to $MES_{\mathcal{L}}(\mathcal{J})$. We say that a mapping $\mathcal{M}_1$ has fewer changes than a mapping $\mathcal{M}_2$ with respect to a mapping $\mathcal{M}$ if $\mathcal{M} \setminus \mathcal{M}_1 \subset \mathcal{M} \setminus \mathcal{M}_2$ (fewer deletions) or $\mathcal{M} \setminus \mathcal{M}_1 = \mathcal{M} \setminus \mathcal{M}_2$ and $\mathcal{M}_1 \setminus \mathcal{M} \subset \mathcal{M}_2 \setminus \mathcal{M}$ (same deletions and fewer insertions).

Similarly to DMR, we introduce an operator, denoted $\circ_{\mathcal{L}}$, to update an OBDA specification through EMR, and a notion of query entailment for $\mathcal{L}$-EMRs.

We finally note that, differently from DMRs, a *finite* $\mathcal{L}$-EMR does not always exist, even if $\mathcal{M}$ is finite, thus the study of this case is particularly challenging.

Complexity of CQ entailment under DMR and EMR semantics. We focus on TBoxes specified in *DL-Lite*$_R$ [3], the logical counterpart of the W3C standard OWL 2 QL, which is a prominent DL in OBDA, and consider both GAV and GLAV mappings. In these settings we provide the following complexity results.

Theorem 1. *Let $\mathcal{T}'$ be a DL-Lite$_R$ TBox, $\mathcal{S}'$ a simple source schema, $\mathcal{M}$ a finite GLAV mapping, q a BCQ over $\mathcal{T}'$, and D a legal instance for $\mathcal{S}'$. Deciding $(\mathcal{M} \bullet \langle \mathcal{T}', \mathcal{S}' \rangle, D) \models q$ is Π_2^p-complete in combined complexity and in* AC^0 *in data complexity.*

As for the EMR case, we first study the setting with GAV mappings and obtain a result analogous to Theorem 1. Roughly speaking, for this setting we are able to reduce

CQ entailment under EMRs to CQ entailment under DMRs through a construction that is polynomial in the size of $\mathcal{T}$ and $\mathcal{M}$ and independent of D.

Theorem 2. *Let $\mathcal{J} = \langle \mathcal{T}, \mathcal{S}, \mathcal{M} \rangle$ be an OBDA specification, where $\mathcal{T}$ is a DL-Lite$_R$ TBox, $\mathcal{S}$ is a simple source schema, and $\mathcal{M}$ is a finite GAV mapping that is globally consistent for $\langle \mathcal{T}, \mathcal{S} \rangle$. Let $\mathcal{T}'$ be a DL-Lite$_R$ TBox, $\mathcal{S}'$ a simple source schema, D a legal instance for $\mathcal{S}'$, and q a BCQ over $\mathcal{T}'$. Deciding $(\mathcal{J} \circ_{GAV} \langle \mathcal{T}', \mathcal{S}' \rangle, D) \models q$ is Π_2^p-complete in combined complexity and in* AC^0 *in data complexity.*

For GLAV, we give a different reduction of CQ entailment under EMR to CQ entailment under DMR, which allows us to establish the data complexity of the problem, under a condition on the interaction between the negative role inclusions (i.e., disjointness role axioms) of the updated TBox $\mathcal{T}'$ (denoted $\mathit{NR}(\mathcal{T}')$) and the initial specification.

Theorem 3. *Let $\mathcal{J} = \langle \mathcal{T}, \mathcal{S}, \mathcal{M} \rangle$ be an OBDA specification, where $\mathcal{T}$ is a DL-Lite$_R$ TBox, $\mathcal{S}$ is a simple source schema, and $\mathcal{M}$ is a finite GLAV mapping. Let $\mathcal{T}'$ be a DL-Lite$_R$ TBox such that $\mathcal{M}$ is globally consistent for $\langle \mathcal{T} \cup \mathit{NR}(\mathcal{T}'), \mathcal{S} \rangle$, $\mathcal{S}'$ a simple source schema, D a legal instance for $\mathcal{S}'$, and q a BCQ over $\mathcal{T}'$. Deciding $(\mathcal{J} \circ_{GLAV} \langle \mathcal{T}', \mathcal{S}' \rangle, D) \models q$ is in* AC^0 *in data complexity.*

References

1. N. Antonioli, F. Castanò, S. Coletta, S. Grossi, D. Lembo, M. Lenzerini, A. Poggi, E. Virardi, and P. Castracane. Ontology-based data management for the italian public debt. In *Proc. of FOIS*, pages 372–385, 2014.
2. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2nd edition, 2007.
3. D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *JAR*, 39(3):385–429, 2007.
4. M. Giese, A. Soylu, G. Vega-Gorgojo, A. Waaler, P. Haase, E. Jiménez-Ruiz, D. Lanti, M. Rezk, G. Xiao, Ö. L. Özçep, and R. Rosati. Optique: Zooming in on big data. *IEEE Computer*, 48(3):60–67, 2015.
5. E. Jiménez-Ruiz, C. Meilicke, B. C. Grau, and I. Horrocks. Evaluating mapping repair systems with large biomedical ontologies. In *Proc. of DL*, volume 1014 of *CEUR*, `ceur-ws.org`, pages 246–257, 2013.
6. D. Lembo, J. Mora, R. Rosati, D. F. Savo, and E. Thorstensen. Mapping analysis in ontology-based data access: Algorithms and complexity. In *Proc. of ISWC*, pages 217–234, 2015.
7. D. Lembo, R. Rosati, V. Santarelli, D. Fabio Savo, and E. Thorstensen. Mapping repair in ontology-based data access evolving systems. In *Proc. of IJCAI*, pages 1160–1166, 2017.
8. C. Meilicke, H. Stuckenschmidt, and A. Tamilin. Reasoning support for mapping revision. *JLC*, 19(5):807–829, 2009.
9. A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Linking data to ontologies. *J. on Data Semantics*, X:133–173, 2008.
10. E. Rahm and P. A. Bernstein. An online bibliography on schema evolution. *SIGMOD Record*, 35(4):30–31, 2006.
11. F. Zablith, G. Antoniou, M. d'Aquin, G. Flouris, H. Kondylakis, E. Motta, D. Plexousakis, and M. Sabou. Ontology evolution: a process-centric survey. *Knowledge Eng. Review*, 30(1):45–75, 2015.