Towards General Representability in Knowledge Exchange

Marcelo Arenas[1], Jorge Pérez[2], and Emanuel Sallinger[3]

[1] Pontificia Universidad Católica de Chile & University of Oxford
[2] Department of Computer Science, Universidad de Chile
[3] Vienna University of Technology

1 Introduction

In data exchange, one is typically given a source database instance and a mapping between a source schema and a target schema. The goal then is to materialize a target database instance that corresponds to the source instance and the mapping. In this setting source data is explicitly given, that is, every fact that is true in it is explicitly mentioned in the source database instance.

Knowledge bases typically consist of some explicit knowledge – like in database instances – and some implicit knowledge, usually given in the form of rules that specify how to derive knowledge not explicitly stored. In [5], Arenas et al. introduced the setting of knowledge exchange, where one is given a source knowledge base and a mapping between a source schema and a target schema, and the goal is to materialize a target knowledge base, that is, both explicit and implicit knowledge.

While the question of what constitutes a "good" solution to a data exchange problem has been the topic of research in recent years [11], the question of what constitutes a "good" solution to a knowledge exchange problem is rather new. In particular, there are now two components, explicit and implicit knowledge, which play rather different roles in what we expect from them.

In [5], two especially desirable properties of the target knowledge base were identified. First, one generally wants to minimize explicit knowledge, thus generating as much knowledge by rules as possible. Second, in typical knowledge-based systems (such as those based on RDFS, OWL, or general description logics), the explicit knowledge changes frequently, but the implicit knowledge remains constant over a longer time. Hence, it is desirable to let the target implicit knowledge only depend on the implicit knowledge at the source and the mapping between source and target. Thus, knowledge exchange effectively becomes a two-stage process:

1. materialize target *implicit* knowledge, given only the source implicit knowledge and the source-to-target mapping; and
2. materialize target *explicit* knowledge w.r.t. the materialized implicit knowledge, which should be as small as possible.

Here, we focus on the former, and thus say that such target implicit knowledge "represents" the source implicit knowledge under a given mapping. Two such notions of "representation" were introduced recently: *safety* [5] and $\mathcal{Q}$-*representability* [3, 4], where $\mathcal{Q}$ is a query language. In this work, we introduce the very broad notion of *general representability* that captures safety as a special case and lays the foundation to the study of representability in a broad setting, in particular with the goal of extending our work to $\mathcal{Q}$-representability. To achieve this, general representability is not limited to knowledge exchange, but based on arbitrary schema mappings. As an essential tool for this study, we develop conditions on schema mapping languages under which notions of equivalence important to our investigation coincide.

Organization. In Section 2, we define the necessary preliminaries. We start introducing general representability in Section 3. Then, in Section 4, we show how to capture safety. In Section 5, we introduce notions of equivalence towards capturing $\mathcal{Q}$-representability. We conclude in Section 6.

2 Preliminaries

We assume some familiarity with database theory and data exchange (cf., e.g., [1, 6]). A schema $\mathbf{S}$ is a set of relation symbols, each having a fixed arity. An instance $I \in \mathsf{Inst}(\mathbf{S})$ assigns to each relation symbol R a relation R^I of the corresponding arity. We denote by $\mathsf{dom}(I)$ the set of all constants that occur in a relation of I. We assume instances to be finite and, thus, $\mathsf{dom}(I)$ is finite.

A *schema mapping* $\mathcal{M}$ (or just *mapping*) from a source schema $\mathbf{S}_1$ to a target schema $\mathbf{S}_2$ is a subset of $\mathsf{Inst}(\mathbf{S}_1) \times \mathsf{Inst}(\mathbf{S}_2)$.[4] The composition of two mappings, symbolized by $\circ$, is defined as the composition of binary relations [9]. $\mathcal{M}$ and $\mathcal{M}'$ are equivalent, written $\mathcal{M} \equiv \mathcal{M}'$, if they are equal as binary relations ($\mathcal{M} = \mathcal{M}'$). Besides, we write $\mathcal{M} \vDash \mathcal{M}'$ for $\mathcal{M} \subseteq \mathcal{M}'$. We later use the notion of a mapping $\mathcal{M}$ "based on" some logical formalism. By that we mean that there exists a set Σ of dependencies from that formalism s.t. $\mathcal{M} = \{(I, J) \mid (I, J) \vDash \Sigma\}$. As these specific formalisms are not the focus of this paper, we refer to e.g. [5] for formal details and definitions. We say that Σ is over a schema $\mathbf{S}$ if all relation symbols occurring in Σ are from $\mathbf{S}$. For a set of instances $\mathcal{I} \subseteq \mathsf{Inst}(\mathbf{S})$, the set of solutions of $\mathcal{I}$ under $\mathcal{M}$ is $\mathsf{Sol}_{\mathcal{M}}(\mathcal{I}) = \{J \mid (I, J) \in \mathcal{M} \text{ and } I \in \mathcal{I}\}$. A single instance $I \in \mathsf{Inst}(\mathbf{S}_1)$ is treated as a singleton set $\{I\}$ wherever applicable, e.g. $\mathsf{Sol}_{\mathcal{M}}(I) = \mathsf{Sol}_{\mathcal{M}}(\{I\})$.

In this paper we deal with knowledge bases in a very general way treating them as mappings over a single schema. A general knowledge base over a schema $\mathbf{S}$ is represented simply as a mapping $\mathcal{M}$ from $\mathbf{S}$ to $\mathbf{S}$ such that, intuitively, if I represents the explicit knowledge, then $\mathsf{Sol}_{\mathcal{M}}(I)$ represents all possible *completions* of I where the implicit knowledge has been made explicit. For keeping schema specifications brief, a subscripted mapping $\mathcal{M}_i$ is always from schema $\mathbf{S}_i$ to $\mathbf{S}_i$, and a mapping $\mathcal{M}_{ij}$ is always from schema $\mathbf{S}_i$ to $\mathbf{S}_j$.

---

[4] To adapt to the knowledge exchange setting [5], we do not explicitly consider labeled nulls in target instances, however the *concept* is preserved as constants only occurring in a target instance are essentially treated as labeled nulls in data exchange [6].

3 General Representability

Intuitively, given mappings $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_{12}$, we say that $\mathcal{M}_2$ "represents" $\mathcal{M}_1$ under $\mathcal{M}_{12}$ if the following equivalence is fulfilled:

$$\mathcal{M}_1 \circ \mathcal{M}_{12} \equiv \mathcal{M}_{12} \circ \mathcal{M}_2. \tag{†}$$

In the context of knowledge exchange, $\mathcal{M}_1$ and $\mathcal{M}_2$ can be seen as the mappings responsible for generating the completions discussed in the preliminaries, and $\mathcal{M}_{12}$ as the mapping used for transferring data. The condition (†) is very strong. In particular, $\mathcal{M}_{12}$ must already transfer enough of the source data for such an $\mathcal{M}_2$ to exist. A very critical point is that we require "logical" equivalence, i.e, for every source instance *all* the solutions must coincide. It can be demonstrated in simple examples that this requirement is too strong to be widely applicable.

Less strict notions of representability have been studied in recent research, namely *safety* [5] in the context of full tuple-generating dependencies [8], and $\mathcal{Q}$-*representability* – where $\mathcal{Q}$ stands for a class of queries – in the context of description logics [3, 4]. We therefore introduce *general representability*, a very weak notion of representability which allows us to study relaxed notions of representability in a very general setting.

Definition 1. *Let $\equiv_\chi$ be an equivalence relation on schema mappings. If there exist mappings $\mathcal{M}_1^\star$ and $\mathcal{M}_{12}^\star$ such that*

$$\mathcal{M}_1 \circ \mathcal{M}_{12} \ \equiv_\chi \ \mathcal{M}_1^\star \circ \mathcal{M}_{12}^\star \circ \mathcal{M}_2, \tag{‡}$$

then $\mathcal{M}_2$ is called a general χ-representation *of $\mathcal{M}_1$ under $\mathcal{M}_{12}$ witnessed by the pair of mappings $(\mathcal{M}_1^\star, \mathcal{M}_{12}^\star)$.*

The intuition behind this definition is the following: $\mathcal{M}_2$ may not be the "ideal" representation as defined by equation (†). Though, if through keeping small parts of $\mathcal{M}_1$ and by slightly extending $\mathcal{M}_{12}$ we can show equivalence, this may be good enough for a given application. In particular, it might not be possible to completely remove $\mathcal{M}_1$, but we may have to keep a "part" of $\mathcal{M}_1$, namely $\mathcal{M}_1^\star$. Second, it might not be possible to use $\mathcal{M}_{12}$, but we may need a slightly "extended" mapping $\mathcal{M}_{12}^\star$. Of course, to study representability in a general setting, we do not formally constrain in this definition how these "parts" and "extensions" must look like. The third way we may adapt our equation to some application is by using an appropriate equivalence relation $\equiv_\chi$, e.g., by using the so-called data exchange equivalence [7] if our application is related to data exchange.

Given this general definition, one has a lot of freedom in choosing $\mathcal{M}_1^\star$ and $\mathcal{M}_{12}^\star$. In particular, by choosing $\mathcal{M}_1^\star = \mathcal{M}_1$ and $\mathcal{M}_{12}^\star = \mathcal{M}_{12}$, even the identity mapping $\mathcal{M}_2 = \mathsf{Id}_2 = \{(I, I) \mid I \in \mathsf{Inst}(\mathbf{S}_2)\}$ is a general representation of $\mathcal{M}_1$ under $\mathcal{M}_{12}$. On the other end of the spectrum, there is a clear "best" kind of general representation $\mathcal{M}_2$, namely if one can choose $\mathcal{M}_1^\star = \mathsf{Id}_1 = \{(I, I) \mid I \in \mathsf{Inst}(\mathbf{S}_1)\}$ and $\mathcal{M}_{12}^\star = \mathcal{M}_{12}$. If these two equalities hold, we speak of a *strong χ-representation*. We thus described what one could consider "trivial" and "best" representations, however there is much room between those two extremes.

For example, it is reasonable to require that while $\mathcal{M}_1^\star \circ \mathcal{M}_{12}^\star$ should be able to be more restrictive than $\mathcal{M}_{12}$, it should not be more restrictive than $\mathcal{M}_1 \circ \mathcal{M}_{12}$ together. Altogether, one can thus require that:

$$\mathcal{M}_1 \circ \mathcal{M}_{12} \vDash \mathcal{M}_1^\star \circ \mathcal{M}_{12}^\star \quad \text{and} \quad \mathcal{M}_1^\star \circ \mathcal{M}_{12}^\star \vDash \mathcal{M}_{12}.$$

When these conditions hold, we talk about a *weak χ-representation*. We intend to use weak χ-representations for characterizing $\mathcal{Q}$-representability in future work.

The intuition behind the notion of weak representability is to offer an intermediate amount of freedom between strong representations and unconstrained general representations. In particular, the witness $(\mathcal{M}_1^\star, \mathcal{M}_{12}^\star)$ has to be at least as restrictive as the witness of a strong representation ($\mathcal{M}_1^\star \circ \mathcal{M}_{12}^\star \vDash \mathcal{M}_{12}$) and at most as restrictive as the witness of a "trivial" representation ($\mathcal{M}_1 \circ \mathcal{M}_{12} \vDash \mathcal{M}_1^\star \circ \mathcal{M}_{12}^\star$). These conditions allow to rule out problematic behaviour that may occur in a general setting, but not in the context of e.g. $\mathcal{Q}$-representability.

An orthogonal restriction is to allow either only a source-to-target witness $\mathcal{M}_{12}^\star$ or only a source witness $\mathcal{M}_1^\star$ in Definition 1, giving rise to the notions of *source-to-target χ-representation* ($\mathcal{M}_1^\star = \mathsf{Id}_1$) and *source χ-representation* ($\mathcal{M}_{12}^\star = \mathcal{M}_{12}$). Formally, $\mathcal{M}_2$ is a *source-to-target χ-representation* of $\mathcal{M}_1$ under $\mathcal{M}_{12}$ if $\mathcal{M}_2$ is a general χ-representation witnessed by the pair of mappings $(\mathsf{Id}_1, \mathcal{M}_{12}^\star)$. Source χ-representation can be defined similarly by requiring $(\mathcal{M}_1^\star, \mathcal{M}_{12})$ as the witness in Definition 1. Note that while for some situations $\mathcal{M}_1^\star$ or $\mathcal{M}_{12}^\star$ suffice, the interplay between them can become important, in particular when putting restrictions on the classes of mappings used. In the next section we use source-to-target χ-representations to characterize safety.

4 Characterizing Safety

In this section, we show that the notion of *safety*, which was developed in [5] in the context of knowledge exchange, can be characterized by the notion of general representation introduced in the previous section. In what follows, we assume that a *knowledge base* $K = (I, \Sigma)$ over a schema $\mathbf{S}$ is given by an instance I of $\mathbf{S}$, called the explicit knowledge, and a set Σ of dependencies over $\mathbf{S}$, called the implicit knowledge. Moreover, we associate to Σ a mapping $\mathsf{Map}(\Sigma) = \{(I, J) \mid I, J \in \mathsf{Inst}(\mathbf{S}), I \subseteq J \text{ and } J \vDash \Sigma\}$. Finally, we define the set of models of $K = (I, \Sigma)$ as $\mathsf{Mod}(K) = \mathsf{Sol}_{\mathsf{Map}(\Sigma)}(I)$.

In order to define the notion of safety proposed in [5], we need to introduce some terminology. Given a set $\mathcal{I}$ of instances of a schema $\mathbf{S}$, we define $\mathsf{Min}(\mathcal{I})$ as the set of minimal instances in $\mathcal{I}$ under the subset relation, that is, as $\{I \in \mathcal{I} \mid \text{there is no } I' \in \mathcal{I} \text{ such that } I' \subsetneq I\}$. Moreover, given a mapping $\mathcal{M}$ from a schema $\mathbf{S}_1$ to a schema $\mathbf{S}_2$ and knowledge bases $K_1 = (I_1, \Sigma_1)$, $K_2 = (I_2, \Sigma_2)$ over $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively, K_2 is called a *minimal knowledge-base solution* for K_1 under $\mathcal{M}$ if $\mathsf{Min}(\mathsf{Mod}(K_2)) = \mathsf{Min}(\mathsf{Sol}_\mathcal{M}(\mathsf{Mod}(K_1)))$ [5]. With these concepts, we can formally define the notion of safety.

Definition 2 ([5]). *Let $\mathcal{M}$ be a schema mapping from a schema $\mathbf{S}_1$ to a schema $\mathbf{S}_2$, and Σ_1, Σ_2 be sets of dependencies over $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively. Then Σ_2 is said to be* safe *for Σ_1 under $\mathcal{M}$, if for every instance I_1 of $\mathbf{S}_1$, there exists an instance I_2 of $\mathbf{S}_2$ such that (I_2, Σ_2) is a minimal knowledge-base solution for (I_1, Σ_1) under $\mathcal{M}$.*

In order to capture safety by using the notion of general representability, we first need to choose the appropriate notion of equivalence to be used in equation (‡). It is clear that the notion of minimality plays a major role in Definition 2, thus equivalence with

respect to minimal solutions is a natural choice. More precisely, for $I \in \mathsf{Inst}(\mathbf{S}_1)$, we define $\mathsf{MinSol}_{\mathcal{M}}(I) = \mathsf{Min}(\mathsf{Sol}_{\mathcal{M}}(I))$ and say that $\mathcal{M}_{12} \equiv_{\mathsf{MinSol}} \mathcal{M}'_{12}$ if for every $I \in \mathsf{Inst}(\mathbf{S}_1)$, it holds that $\mathsf{MinSol}_{\mathcal{M}_{12}}(I) = \mathsf{MinSol}_{\mathcal{M}'_{12}}(I)$.

It is important to notice that the original definition of safety given in [5] only considered full tuple-generating dependencies (that is, Σ_1 and $\mathcal{M}$ in Definition 2 were required to be based on these dependencies). A consequence of this restriction is that only mappings that are *absolutely consistent*, in the sense defined in [2], were considered in the original definition of safety. So when capturing this notion we focus on this class of mappings. More precisely, a mapping $\mathcal{M}$ from a schema $\mathbf{S}_1$ to a schema $\mathbf{S}_2$ is said to be absolutely consistent if for every $I \in \mathsf{Inst}(\mathbf{S}_1)$, it holds that $\mathsf{Sol}_{\mathcal{M}}(I) \neq \emptyset$. Then we have that:

Proposition 1. *Let Σ_1, Σ_2 be sets of dependencies over schemas $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively, and assume that $\mathcal{M}_{12}$ is a mapping such that $\mathsf{Map}(\Sigma_1) \circ \mathcal{M}_{12}$ is absolutely consistent. Then Σ_2 is safe for Σ_1 under $\mathcal{M}_{12}$ iff $\mathsf{Map}(\Sigma_2)$ is a source-to-target MinSol-representation of $\mathsf{Map}(\Sigma_1)$ under $\mathcal{M}_{12}$.*

It follows that safety as defined in [5] can be captured by source-to-target MinSol-representability, as $\mathcal{M}_1 \circ \mathcal{M}_{12}$ is absolutely consistent in the definition given in [5] (since $\mathcal{M}_1$ and $\mathcal{M}_{12}$ are required to be based on full tuple-generating dependencies in [5]). Interestingly, the previous result also holds for mappings based on tuple-generating dependencies with terminating chase.

5   On the Equivalence of Mappings

To capture notions like safety or $\mathcal{Q}$-representability within the framework of general representability, one of the key ingredients is choosing the right notion of equivalence of mappings. In this section, we present the first steps towards the development of some appropriate notions of equivalence of mappings, which connect minimal solutions (used to characterize safety in Proposition 1) to notions of equivalence relevant for $\mathcal{Q}$-representability.

A key role in this connection is played by a generalization of the widely used notion of universal solution [6], which is defined next. Let $\mathcal{I}$ be a set of instances, $I_1, I_2 \in \mathcal{I}$, and let $\mathcal{C}$ be a set of constants. A homomorphism from I_1 to I_2 that is constant on $\mathcal{C}$ is a function $h : \mathsf{dom}(I_1) \to \mathsf{dom}(I_2)$ such that for all $c \in (\mathcal{C} \cap \mathsf{dom}(I_1))$ we have that $h(c) = c$, and for all $(x_1, \ldots, x_n) \in R^{I_1}$ we have that $(h(x_1), \ldots, h(x_n)) \in R^{I_2}$. We write $I_1 \preceq_{\mathcal{C}} I_2$ if such an h exists. We say that $I_1 \prec_{\mathcal{C}} I_2$ if $I_1 \preceq_{\mathcal{C}} I_2$ and $I_2 \npreceq_{\mathcal{C}} I_1$. Then let $\mathsf{Hom}_{\mathcal{C}}(\mathcal{I}) = \{I \in \mathcal{I} \mid$ there is no $I' \in \mathcal{I}$ such that $I' \prec_{\mathcal{C}} I\}$. Universal solutions are defined as the minimum elements of a set $\mathcal{I}$ of solutions up to equivalence under homomorphisms, while the elements of $\mathsf{Hom}_{\mathcal{C}}(\mathcal{I})$ are defined as the minimal elements of $\mathcal{I}$ up to equivalence under homomorphisms. Furthermore, given a pair I_1, I_2 of instances of a schema $\mathbf{S}$ and a set $\mathcal{C}$ of constants, we say that I_1 is a $\mathcal{C}$-core of I_2 if [10, 8]: (1) I_1 is a sub-instance of I_2, (2) $I_2 \preceq_{\mathcal{C}} I_1$, and (3) there is no homomorphism from I_2 to a proper sub-instance of I_1 that is the identity on $\mathcal{C}$. Then let $\mathsf{Core}_{\mathcal{C}}(\mathcal{I}) = \{I \in \mathsf{Inst}(\mathbf{S}) \mid I$ is a $\mathcal{C}$-core of some $I' \in \mathcal{I}\}$.

We are now ready to define the notions of equivalence of mappings studied in this section. For $I \in \mathsf{Inst}(\mathbf{S}_1)$ and mapping $\mathcal{M}$ from $\mathbf{S}_1$, we define $\mathsf{HomSol}_{\mathcal{M}}(I) =$

$\mathsf{Hom}_{\mathrm{dom}(I)}(\mathsf{Sol}_{\mathcal{M}}(I))$ and $\mathsf{CoreHomSol}_{\mathcal{M}}(I) = \mathsf{Core}_{\mathrm{dom}(I)}(\mathsf{HomSol}_{\mathcal{M}}(I)))$. More-over, $\mathcal{M}_{12} \equiv_{\mathsf{CoreHomSol}} \mathcal{M}'_{12}$ if for all $I \in \mathsf{Inst}(\mathbf{S}_1)$ we have that $\mathsf{CoreHomSol}_{\mathcal{M}_{12}}(I) = \mathsf{CoreHomSol}_{\mathcal{M}'_{12}}(I)$. Next we introduce conditions on mappings that allow us to relate the newly introduced notion to the notion of MinSol we used before.

Definition 3. *Let $\mathcal{M}$ be a mapping from a schema $\mathbf{S}_1$ to a schema $\mathbf{S}_2$. Then: $\mathcal{M}$ is* domain-full *if for all $I \in \mathsf{Inst}(\mathbf{S}_1)$ and $J \in \mathsf{HomSol}_{\mathcal{M}}(I)$, there exists $J' \in \mathsf{HomSol}_{\mathcal{M}}(I)$ such that $J' \subseteq J$ and $\mathrm{dom}(J') \subseteq \mathrm{dom}(I)$; $\mathcal{M}$ is* founded *if for all $I \in \mathsf{Inst}(\mathbf{S}_1)$ and $J \in \mathsf{Sol}_{\mathcal{M}}(I)$, there exists $J' \in \mathsf{HomSol}_{\mathcal{M}}(I)$ such that $J' \preceq_{\mathrm{dom}(I)} J$; and $\mathcal{M}$ admits core-solutions* if for all $I \in \mathsf{Inst}(\mathbf{S}_1)$ we have $\mathsf{CoreHomSol}_{\mathcal{M}}(I) \subseteq \mathsf{Sol}_{\mathcal{M}}(I)$.

Proposition 2. *Let $\mathcal{M}$ and $\mathcal{M}'$ be mappings from a schema $\mathbf{S}_1$ to a schema $\mathbf{S}_2$ such that $\mathcal{M}$ is domain-full, founded, and admits core-solutions. Then it holds that $\mathcal{M} \equiv_{\mathsf{MinSol}} \mathcal{M}'$ iff $\mathcal{M} \equiv_{\mathsf{CoreHomSol}} \mathcal{M}'$.*

So for safety as originally defined in [5], we thus know that the following holds:

Corollary 1. *Let Σ_1 be a set of full tgds over a schema $\mathbf{S}_1$, Σ_2 a set of dependencies over a schema $\mathbf{S}_2$ and $\mathcal{M}_{12}$ a mapping from $\mathbf{S}_1$ to $\mathbf{S}_2$ based on full s-t tgds. Then Σ_2 is safe for Σ_1 under $\mathcal{M}_{12}$ iff $\mathsf{Map}(\Sigma_2)$ is a source-to-target $\mathsf{CoreHomSol}$-representation of $\mathsf{Map}(\Sigma_1)$ under $\mathcal{M}_{12}$.*

We have not yet developed a similar result for the notion of $\mathcal{Q}$-representability. But towards that, we studied the connection between equivalence based on CoreHomSol and notions based on query languages relevant for $\mathcal{Q}$-representability, that we introduce next. We assume that a k-ary *query* q over a schema $\mathbf{S}$ is a function that maps every instance I of $\mathbf{S}$ to a k-ary relation $q(I)$ contained in $\mathrm{dom}(I)^k$. Moreover, given a set $\mathcal{I}$ of instances of $\mathbf{S}$ and a k-ary query q over $\mathbf{S}$, the set of certain answers of q over $\mathcal{I}$, denoted by $\mathsf{cert}_q(\mathcal{I})$, is defined as $\bigcap_{I \in \mathcal{I}} q(I)$. Let $\mathcal{Q}$ be a class of queries, and $\mathcal{M}, \mathcal{M}'$ be mappings from a schema $\mathbf{S}_1$ to a schema $\mathbf{S}_2$. In the spirit of [7], we say that $\mathcal{M} \equiv_{\mathcal{Q}} \mathcal{M}'$ if for all $q \in \mathcal{Q}$ over $\mathbf{S}_2$ and $I \in \mathsf{Inst}(\mathbf{S}_1)$, it holds $\mathsf{cert}_q(\mathsf{Sol}_{\mathcal{M}}(I)) = \mathsf{cert}_q(\mathsf{Sol}_{\mathcal{M}'}(I))$. A k-ary query q over a schema $\mathbf{S}$ is said to be closed under homomorphisms if for every pair of instances I_1, I_2 of $\mathbf{S}$, tuple $(a_1, \dots, a_k) \in q(I_1)$ and homomorphism h from I_1 to I_2, it holds that $(h(a_1), \dots, h(a_k)) \in q(I_2)$. Let CHQ be the class of queries that are closed under homomorphisms, and UCQ be the class of unions of conjunctive queries.

Again, to formally state the relations between notions of equivalence of mappings, we define some natural conditions and the necessary terminology. We use notation $I_1 \cong_{\mathcal{C}} I_2$ to indicate that there exists an isomorphism from I_1 to I_2 that is the identity on $\mathcal{C}$, and we use notation $I_1 \leftrightarrow_{\mathcal{C}} I_2$ to indicate that $I_1 \preceq_{\mathcal{C}} I_2$ and $I_2 \preceq_{\mathcal{C}} I_1$. Besides, we use notation $I_1 \cong I_2$ as a shorthand for $I_1 \cong_{\emptyset} I_2$.

Definition 4. *Let $\mathcal{M}$ be a mapping from a schema $\mathbf{S}_1$ to a schema $\mathbf{S}_2$. Then: $\mathcal{M}$ is* closed under isomorphisms *if for all $I_1, I_2 \in \mathsf{Inst}(\mathbf{S}_1)$ and $J_1, J_2 \in \mathsf{Inst}(\mathbf{S}_2)$, if $J_1 \in \mathsf{Sol}_{\mathcal{M}}(I_1)$ and $(I_1, J_1) \cong (I_2, J_2)$, then $J_2 \in \mathsf{Sol}_{\mathcal{M}}(I_2)$; $\mathcal{M}$ is* target domain-independent *if for all $I \in \mathsf{Inst}(\mathbf{S}_1)$ and $J_1, J_2 \in \mathsf{HomSol}_{\mathcal{M}}(I)$, it holds that $\mathrm{dom}(J_1) \cap \mathrm{dom}(I) = \mathrm{dom}(J_2) \cap \mathrm{dom}(I)$; and $\mathcal{M}$ has a finite base* if for every $I \in \mathsf{Inst}(\mathbf{S}_1)$, it holds that $\mathsf{HomSol}_{\mathcal{M}}(I)$ is finite up to $\leftrightarrow_{\mathrm{dom}(I)}$.

Theorem 1. *Let $\mathcal{M}$, $\mathcal{M}'$ be mappings from a schema $\mathbf{S}_1$ to a schema $\mathbf{S}_2$, such that $\mathcal{M}$, $\mathcal{M}'$ are closed under isomorphisms, founded and target domain-independent. Then $\mathcal{M} \equiv_{\mathsf{CHQ}} \mathcal{M}'$ iff $\mathcal{M} \equiv_{\mathsf{CoreHomSol}} \mathcal{M}'$*

As a corollary of the proof of Theorem 1, we have the following for $\equiv_{\mathsf{UCQ}}$, the most studied notion of equivalence in the context of $\mathcal{Q}$-representability:

Corollary 2. *Let $\mathcal{M}$, $\mathcal{M}'$ be mappings from a schema $\mathbf{S}_1$ to a schema $\mathbf{S}_2$, such that $\mathcal{M}$, $\mathcal{M}'$ are closed under isomorphisms, are founded, are target domain-independent and have a finite base. Then $\mathcal{M} \equiv_{\mathsf{UCQ}} \mathcal{M}'$ iff $\mathcal{M} \equiv_{\mathsf{CoreHomSol}} \mathcal{M}'$.*

6 Concluding remarks

In this paper, we introduced the framework of general representability that allows us to study notions of representability like safety and $\mathcal{Q}$-representability. We introduced notions of equivalence suitable for capturing safety, and defined conditions under which these characterizations apply. In particular, we were able to capture safety as originally defined. We also started introducing the necessary machinery for studying $\mathcal{Q}$-representability, with a focus on the UCQ and the more general CHQ case.

As future work, we intend to capture $\mathcal{Q}$-representability, and continue using the framework of general representability to develop notions of representation both for specific applications like knowledge exchange as well as in the general case. The study of the conditions and notions of equivalence we introduced will also be an interesting endeavor in its own right, as it allows us to find candidates for conditions which schema mapping formalisms should satisfy.

References

1. Abiteboul, S., Hull, R., Vianu, V.: Foundations of Databases. Addison-Wesley (1995)
2. Amano, S., Libkin, L., Murlak, F.: XML schema mappings. In: PODS. pp. 33–42 (2009)
3. Arenas, M., Botoeva, E., Calvanese, D.: Knowledge base exchange. In: DL (2011)
4. Arenas, M., Botoeva, E., Calvanese, D., Ryzhikov, V., Sherkhonov, E.: Exchanging description logic knowledge bases. In: KR (2012)
5. Arenas, M., Pérez, J., Reutter, J.L.: Data exchange beyond complete data. In: PODS. pp. 83–94 (2011)
6. Fagin, R., Kolaitis, P.G., Miller, R.J., Popa, L.: Data exchange: semantics and query answering. Theor. Comput. Sci. 336(1), 89–124 (2005)
7. Fagin, R., Kolaitis, P.G., Nash, A., Popa, L.: Towards a theory of schema-mapping optimization. In: PODS. pp. 33–42 (2008)
8. Fagin, R., Kolaitis, P.G., Popa, L.: Data exchange: getting to the core. ACM Trans. Database Syst. 30(1), 174–210 (2005)
9. Fagin, R., Kolaitis, P.G., Popa, L., Tan, W.C.: Composing schema mappings: Second-order dependencies to the rescue. ACM Trans. Database Syst. 30(4), 994–1055 (2005)
10. Hell, P., Nesetril, J.: The core of a graph. Discrete Mathematics 109(1–3), 117–126 (1992)
11. Hernich, A., Schweikardt, N.: Logic and data exchange: Which solutions are "good" solutions? In: LOFT. pp. 61–85 (2008)