Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management

legacy DB

client app
Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management

The new DB should only contain the information transferred by $\mathcal{M}$. 

Extract operation
Information and redundancy: fundamental concepts in schema mapping management

The new DB should only contain the information transferred by $\mathcal{M}$. 

Extract operation
Information and redundancy: fundamental concepts in schema mapping management

The new DB should only contain the information transferred by $\mathcal{M}$. 

Extract operation
Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management

DB₁  DB₂
Information and redundancy:
fundamental concepts in schema mapping management

DB₁

DB₂
**Information and redundancy:**
fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management

\[ M_1 \rightarrow M \rightarrow M_2 \]

DB$_1$ \hspace{2cm} \[ M \] \hspace{2cm} DB$_2$

\[ M_1 \rightarrow \text{consolidated DB} \rightarrow M_2 \]

Merge operation
Information and redundancy: fundamental concepts in schema mapping management

The new DB should only store the non redundant information w.r.t. $\mathcal{M}$. 

**Diagram:**
- $\mathcal{M}_1$ links $DB_1$ to the consolidated DB.
- $\mathcal{M}_2$ links the consolidated DB to $DB_2$.
- $\mathcal{M}$ links $DB_1$ to $DB_2$. 

**Legend:**
- $DB_1$: Database 1
- $DB_2$: Database 2
- $\mathcal{M}_1$: Map 1
- $\mathcal{M}_2$: Map 2
- $\mathcal{M}$: Merge operation
- **consolidated DB**: The new database that stores non redundant information.
**Information and redundancy:**

fundamental concepts in schema mapping management

The new DB should only store the *non redundant information* w.r.t. $M$. 
Information and redundancy:
fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management
Information and redundancy:
fundamental concepts in schema mapping management
Information and redundancy: fundamental concepts in schema mapping management

Inverse operation
Information and redundancy: fundamental concepts in schema mapping management

Inverse operation

Invertibility for $\mathcal{M}$ should coincide with no loss of information.
Information and redundancy:
fundamental concepts in schema mapping management

Inverse operation

Invertibility for $\mathcal{M}$ should coincide with no loss of information.
Information and redundancy: fundamental concepts in schema mapping management

Inverse operation

Although fundamental, the notions of information and redundancy have received little attention in the schema mapping context.
Foundations of Schema Mapping Management

Marcelo Arenas, Jorge Pérez, Juan L. Reutter, Cristian Riveros

PUC Chile, U. Edinburgh, U. Oxford
We provide foundations for schema mapping management by formalizing the notions of *information* and *redundancy*. 
We provide foundations for schema mapping management by formalizing the notions of *information* and *redundancy*.

Main contributions:

1. *Information* and *redundancy* in schema mappings
   - general formalization
   - characterizations and algorithmic issues
We provide foundations for schema mapping management by formalizing the notions of information and redundancy.

Main contributions:

1. Information and redundancy in schema mappings
   - general formalization
   - characterizations and algorithmic issues

2. Applications of the notions:
   - schema evolution problem
   - Extract, Merge and Inverse operators
Outline

Motivation

Source information
  Algorithmic issues
  Application: Invertibility

Target information
  Application: Extract, first approach

Target and source redundancy
  Application: Extract

Concluding remarks
A bit of notation...

A mapping $\mathcal{M}$ is a set of pairs $(I, J)$ with

- $I$ a source instance and $J$ a target instance
  ($J$ is called a solution for $I$ under $\mathcal{M}$).
A bit of notation...

A *mapping* $\mathcal{M}$ is a set of pairs $(I, J)$ with

- $I$ a source instance and $J$ a target instance
  ($J$ is called a *solution* for $I$ under $\mathcal{M}$).
- the *composition of mappings*, $\mathcal{M} \circ \mathcal{M}'$, is the usual composition of binary relations.
A bit of notation...

A mapping $\mathcal{M}$ is a set of pairs $(I, J)$ with
- $I$ a source instance and $J$ a target instance ($J$ is called a solution for $I$ under $\mathcal{M}$).
- the composition of mappings, $\mathcal{M} \circ \mathcal{M}'$, is the usual composition of binary relations.

Mappings can be specified by formulas (dependencies):

$$\varphi_S(\bar{x}) \rightarrow \psi_T(\bar{x})$$
A bit of notation...

A *mapping* $\mathcal{M}$ is a set of pairs $(I, J)$ with

- $I$ a source instance and $J$ a target instance ($J$ is called a *solution* for $I$ under $\mathcal{M}$).

- the *composition of mappings*, $\mathcal{M} \circ \mathcal{M}'$, is the usual composition of binary relations.

Mappings can be specified by formulas (dependencies):

$$\varphi_S(\bar{x}) \rightarrow \psi_T(\bar{x})$$

with $\varphi_S(\bar{x})$ formula over the source and $\psi_T(\bar{x})$ over the target.
A bit of notation...

A **mapping** $\mathcal{M}$ is a set of pairs $(I, J)$ with

- $I$ a source instance and $J$ a target instance ($J$ is called a **solution** for $I$ under $\mathcal{M}$).
- the **composition of mappings**, $\mathcal{M} \circ \mathcal{M}'$, is the usual composition of binary relations.

Mappings can be specified by formulas (dependencies):

$$\varphi_S(\bar{x}) \rightarrow \psi_T(\bar{x})$$

with $\varphi_S(\bar{x})$ formula over the source and $\psi_T(\bar{x})$ over the target.

- **L$_1$-to-L$_2$ dependency**: $\varphi_S(\bar{x}) \in L_1$ and $\psi_T(\bar{x}) \in L_2$. 
A bit of notation...

A mapping \( \mathcal{M} \) is a set of pairs \((I, J)\) with

- \( I \) a source instance and \( J \) a target instance
  (\( J \) is called a solution for \( I \) under \( \mathcal{M} \)).
- the composition of mappings, \( \mathcal{M} \circ \mathcal{M}' \), is the usual composition of binary relations.

Mappings can be specified by formulas (dependencies):

\[
\varphi_S(\bar{x}) \rightarrow \psi_T(\bar{x})
\]

with \( \varphi_S(\bar{x}) \) formula over the source and \( \psi_T(\bar{x}) \) over the target.

- \( L_1\text{-to-}L_2 \) dependency: \( \varphi_S(\bar{x}) \in L_1 \) and \( \psi_T(\bar{x}) \in L_2 \).
- \( \text{CQ-to-CQ} = \text{st-tgds} \).
- we are also interested in \( \text{CQ} \neq \text{-to-CQ} \) and \( \text{FO-to-CQ} \).
Source information transferred by a mapping: Intuition

Source: \{Emp(\text{name}, \text{lives\_in}, \text{works\_in}) \}
Target_1: \{Person(\text{ssn}, \text{name}) \}
Source information transferred by a mapping: Intuition

Source: \{\textsc{Emp(name, lives\_in, works\_in)} \}
Target$_1$: \{\textsc{Person(ssn, name)} \}

\[ \mathcal{M}_1: \quad \textsc{Emp}(x, y, z) \rightarrow \exists u \textsc{Person}(u, x) \]
Source information transferred by a mapping: Intuition

Source: \{\text{Emp(name, lives\_in, works\_in)} \}
Target$_1$: \{\text{Person(ssn, name)} \}
Target$_2$: \{\text{ENames(name), WorksIn(name, place)} \}

\[ M_1: \text{Emp}(x, y, z) \rightarrow \exists u \text{ Person}(u, x) \]
Source information transferred by a mapping: Intuition

Source: \{Emp(name, lives_in, works_in) \}
Target$_1$: \{Person(ssn, name) \}
Target$_2$: \{ENames(name), WorksIn(name, place) \}

\[ M_1: \text{Emp}(x, y, z) \rightarrow \exists u \text{ Person}(u, x) \]
\[ M_2: \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z) \]
Source information transferred by a mapping: Intuition

Source: \{\text{Emp}(\text{name, lives\_in, works\_in}) \}

Target\_1: \{\text{Person}(\text{ssn, name}) \}

Target\_2: \{\text{ENames}(\text{name}), \text{WorksIn}(\text{name, place}) \}

\[ M_1: \text{Emp}(x, y, z) \rightarrow \exists u \text{ Person}(u, x) \]

\[ M_2: \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z) \]

Intuitively:

\[ M_2 \text{ is more source-informative than } M_1. \]
Source information transferred by a mapping: Intuition

Source: \{\text{Emp(name, lives\_in, works\_in)} \}

Target$_1$: \{\text{Person(ssn, name)} \}

Target$_2$: \{\text{ENames(name), WorksIn(name, place)} \}

Target$_3$: \{\text{Workplace(place)} \}

\[ M_1 : \text{Emp}(x, y, z) \rightarrow \exists u \ \text{Person}(u, x) \]

\[ M_2 : \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z) \]

Intuitively:

\[ M_2 \text{ is more source-informative than } M_1. \]
Source information transferred by a mapping: Intuition

Source: \{Emp(name, lives_in, works_in) \}
Target\_1: \{Person(ssn, name) \}
Target\_2: \{ENames(name), WorksIn(name, place) \}
Target\_3: \{Workplace(place) \}

\[ \mathcal{M}_1: \text{Emp}(x, y, z) \rightarrow \exists u \text{ Person}(u, x) \]
\[ \mathcal{M}_2: \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z) \]
\[ \mathcal{M}_3: \text{Emp}(x, y, z) \rightarrow \text{Workplace}(z) \]

Intuitively:

\[ \mathcal{M}_2 \text{ is more source-informative than } \mathcal{M}_1. \]
Source information transferred by a mapping: Intuition

Source: \{\text{Emp(name, lives\_in, works\_in)} \}
Target$_1$: \{\text{Person(ssn, name)} \}
Target$_2$: \{\text{ENames(name), WorksIn(name, place)} \}
Target$_3$: \{\text{Workplace(place)} \}

\[ M_1: \text{Emp}(x, y, z) \rightarrow \exists u \text{ Person}(u, x) \]
\[ M_2: \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z) \]
\[ M_3: \text{Emp}(x, y, z) \rightarrow \text{Workplace}(z) \]

Intuitively:

\[ M_2 \text{ is more source-informative than } M_1. \]
\[ M_2 \text{ is more source-informative than } M_3. \]
Source information transferred by a mapping: Intuition

Source: \{Emp(name, lives_in, works_in) \}
Target_1: \{Person(ssn, name) \}
Target_2: \{ENames(name), WorksIn(name, place) \}
Target_3: \{Workplace(place) \}

\[ M_1: \quad \text{Emp}(x, y, z) \rightarrow \exists u \quad \text{Person}(u, x) \]
\[ M_2: \quad \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z) \]
\[ M_3: \quad \text{Emp}(x, y, z) \rightarrow \text{Workplace}(z) \]

Intuitively:

\[ M_2 \text{ is more source-informative than } M_1. \]
\[ M_2 \text{ is more source-informative than } M_3. \]
\[ M_1 \text{ and } M_3 \text{ are incomparable.} \]
Assume that $\mathcal{M}_1$ and $\mathcal{M}_2$ share the source schema.

**Definition**

$\mathcal{M}_2$ is *more (or equally) source-informative than* $\mathcal{M}_1$, denoted by

$$\mathcal{M}_1 \preceq_s \mathcal{M}_2,$$
Assume that $\mathcal{M}_1$ and $\mathcal{M}_2$ share the source schema.

**Definition**

$\mathcal{M}_2$ is *more (or equally) source-informative than* $\mathcal{M}_1$, denoted by

$$\mathcal{M}_1 \preceq_s \mathcal{M}_2,$$

if there exists a mapping $\mathcal{M}'$ such that $\mathcal{M}_2 \circ \mathcal{M}' = \mathcal{M}_1$. 


Source information transferred by a mapping: Formalization

Assume that $M_1$ and $M_2$ share the source schema.

**Definition**

$M_2$ is *more (or equally) source-informative than* $M_1$, denoted by

$$M_1 \preceq_s M_2,$$

if there exists a mapping $M'$ such that $M_2 \circ M' = M_1$.

$M_2$ transfers information enough to reconstruct $M_1$. 

Assume that $\mathcal{M}_1$ and $\mathcal{M}_2$ share the source schema.

**Definition**

$\mathcal{M}_2$ is *more (or equally) source-informative than* $\mathcal{M}_1$, denoted by $\mathcal{M}_1 \preceq_s \mathcal{M}_2$, if there exists a mapping $\mathcal{M}'$ such that $\mathcal{M}_2 \circ \mathcal{M}' = \mathcal{M}_1$.

$\mathcal{M}_2$ transfers information enough to reconstruct $\mathcal{M}_1$. 
Source information transferred by a mapping: Formalization

Assume that $\mathcal{M}_1$ and $\mathcal{M}_2$ share the source schema.

Definition

$\mathcal{M}_2$ is more (or equally) source-informative than $\mathcal{M}_1$, denoted by $\mathcal{M}_1 \preceq_s \mathcal{M}_2$, if there exists a mapping $\mathcal{M}'$ such that $\mathcal{M}_2 \circ \mathcal{M}' = \mathcal{M}_1$.

$\mathcal{M}_2$ transfers information enough to reconstruct $\mathcal{M}_1$. 

\[ \text{\textbf{\textup{M}}_2 \textbf{t}r\textbf{a}n\textbf{s}f\textbf{e}r\textbf{s} \textbf{i}n\textbf{f}o\textbf{r}m\textbf{a}t\textbf{i}o\textbf{n} \textbf{e}n\textbf{n}o\textbf{u}gh \textbf{t}o \textbf{r}\textbf{e}\textbf{c}o\textbf{n}\textbf{u}\textbf{s}\textbf{t} \textbf{M}_1.} \]
Assume that $\mathcal{M}_1$ and $\mathcal{M}_2$ share the source schema.

**Definition**

$\mathcal{M}_2$ is *more (or equally) source-informative than* $\mathcal{M}_1$, denoted by $\mathcal{M}_1 \preceq_s \mathcal{M}_2$, if there exists a mapping $\mathcal{M}'$ such that $\mathcal{M}_2 \circ \mathcal{M}' = \mathcal{M}_1$.

$\mathcal{M}_2$ transfers information enough to reconstruct $\mathcal{M}_1$.

\[\mathcal{M}_1: \text{Emp}(x, y, z) \rightarrow \exists u \text{ Person}(u, x)\]
Assume that $\mathcal{M}_1$ and $\mathcal{M}_2$ share the source schema.

**Definition**

$\mathcal{M}_2$ is *more (or equally) source-informative than* $\mathcal{M}_1$, denoted by

$$\mathcal{M}_1 \preceq_s \mathcal{M}_2,$$

if there exists a mapping $\mathcal{M}'$ such that $\mathcal{M}_2 \circ \mathcal{M}' = \mathcal{M}_1$.

$\mathcal{M}_2$ transfers information enough to reconstruct $\mathcal{M}_1$.

\[
\begin{align*}
\mathcal{M}_1 : & \quad \text{Emp}(x, y, z) \rightarrow \exists u \text{ Person}(u, x) \\
\mathcal{M}_2 : & \quad \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z)
\end{align*}
\]
Source information transferred by a mapping: Formalization

Assume that $\mathcal{M}_1$ and $\mathcal{M}_2$ share the source schema.

**Definition**

$\mathcal{M}_2$ is *more (or equally) source-informative than* $\mathcal{M}_1$, denoted by

$$\mathcal{M}_1 \preceq_s \mathcal{M}_2,$$

if there exists a mapping $\mathcal{M}'$ such that $\mathcal{M}_2 \circ \mathcal{M}' = \mathcal{M}_1$.

$\mathcal{M}_2$ transfers information enough to reconstruct $\mathcal{M}_1$.

\[\begin{align*}
\mathcal{M}_1 : \text{Emp}(x, y, z) &\rightarrow \exists u \text{ Person}(u, x) \\
\mathcal{M}_2 : \text{Emp}(x, y, z) &\rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z) \\
\mathcal{M}' : \text{ENames}(x) &\rightarrow \exists u \text{ Person}(u, x)
\end{align*}\]
Assume that $M_1$ and $M_2$ share the source schema.

**Definition**

$M_2$ is *more (or equally) source-informative than* $M_1$, denoted by $M_1 \preceq_s M_2$, if there exists a mapping $M'$ such that $M_2 \circ M' = M_1$.

$M_2$ transfers information enough to reconstruct $M_1$.

$$
M_1 : \text{Emp}(x, y, z) \rightarrow \exists u \ \text{Person}(u, x) \\
M_2 : \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z) \\
M' : \text{ENames}(x) \rightarrow \exists u \ \text{Person}(u, x) \\
M_2 \circ M' = M_1 \implies M_1 \preceq_s M_2
$$
Axiomatization of $\preceq_S$

In the paper, we first define 4 *axioms* for an order $\preceq$ on mappings
Axiomatization of $\leq_S$

In the paper, we first define 4 axioms for an order $\leq$ on mappings

(C1) \textit{reflexivity} : $\mathcal{M} \leq \mathcal{M}$

(C2) \textit{transitivity} : $\mathcal{M}_1 \leq \mathcal{M}_2$ and $\mathcal{M}_2 \leq \mathcal{M}_3$, then $\mathcal{M}_1 \leq \mathcal{M}_3$
Axiomatization of $\leq_S$

In the paper, we first define 4 *axioms* for an order $\leq$ on mappings

(C1) *reflexivity* : $\mathcal{M} \leq \mathcal{M}$
(C2) *transitivity* : $\mathcal{M}_1 \leq \mathcal{M}_2$ and $\mathcal{M}_2 \leq \mathcal{M}_3$, then $\mathcal{M}_1 \leq \mathcal{M}_3$
(C3) *maximum* : $\mathcal{M} \leq \text{Id} = \{(I, I) \mid I \text{ is a source instance}\}$
Axiomatization of $\preceq_S$

In the paper, we first define 4 axioms for an order $\preceq$ on mappings

(C1) reflexivity : $\mathcal{M} \preceq \mathcal{M}$

(C2) transitivity : $\mathcal{M}_1 \preceq \mathcal{M}_2$ and $\mathcal{M}_2 \preceq \mathcal{M}_3$, then $\mathcal{M}_1 \preceq \mathcal{M}_3$

(C3) maximum : $\mathcal{M} \preceq \text{Id} = \{(l, l) \mid l \text{ is a source instance}\}$

(C4) preservation : $\mathcal{M}_1 \preceq \mathcal{M}_2$ then $\mathcal{M} \circ \mathcal{M}_1 \preceq \mathcal{M} \circ \mathcal{M}_2$
Axiomatization of $\preceq_S$

In the paper, we first define 4 axioms for an order $\preceq$ on mappings

(C1) **reflexivity** : $\mathcal{M} \preceq \mathcal{M}$

(C2) **transitivity** : $\mathcal{M}_1 \preceq \mathcal{M}_2$ and $\mathcal{M}_2 \preceq \mathcal{M}_3$, then $\mathcal{M}_1 \preceq \mathcal{M}_3$

(C3) **maximum** : $\mathcal{M} \preceq \text{Id} = \{(l, l) \mid l \text{ is a source instance}\}$

(C4) **preservation** : $\mathcal{M}_1 \preceq \mathcal{M}_2$ then $\mathcal{M} \circ \mathcal{M}_1 \preceq \mathcal{M} \circ \mathcal{M}_2$
Axiomatization of $\preceq_S$

In the paper, we first define 4 axioms for an order $\preceq$ on mappings:

(C1) **reflexivity** : $M \preceq M$

(C2) **transitivity** : $M_1 \preceq M_2$ and $M_2 \preceq M_3$, then $M_1 \preceq M_3$

(C3) **maximum** : $M \preceq \text{Id} = \{(I, I) \mid I \text{ is a source instance}\}$

(C4) **preservation** : $M_1 \preceq M_2$ then $M \circ M_1 \preceq M \circ M_2$

Theorem

*The order $\preceq_S$ is the strictest relation that satisfies (C1-C4).*
Towards deciding $\preceq_S$: target rewritability

Certain answers

Mapping $\mathcal{M}$, target query $Q_T$, source instance $I$:

$$\text{certain}_{\mathcal{M}}(Q_T, I) = \bigcap_{(I,J) \in \mathcal{M}} Q_T(J)$$
Towards deciding $\leq_S$: target rewritability

**Certain answers**

Mapping $\mathcal{M}$, target query $Q_T$, source instance $I$:

$$\text{certain}_\mathcal{M}(Q_T, I) = \bigcap_{(I, J) \in \mathcal{M}} Q_T(J)$$

**Definition**

A source query $Q_S$ is *target rewritable under* $\mathcal{M}$ if there exists a target query $Q_T$ such that

$$Q_S(I) = \text{certain}_\mathcal{M}(Q_T, I)$$

for every source instance $I$. 
Towards deciding $\leq_S$: target rewritability

*Certain answers*

Mapping $\mathcal{M}$, target query $Q_T$, source instance $I$:

$$\text{certain}_\mathcal{M}(Q_T, I) = \bigcap_{(I,J) \in \mathcal{M}} Q_T(J)$$

**Definition**

A source query $Q_S$ is *target rewritable under* $\mathcal{M}$ if there exists a target query $Q_T$ such that

$$Q_S(I) = \text{certain}_\mathcal{M}(Q_T, I)$$

for every source instance $I$.

▶ Intuitively: if $Q_S$ is target rewritable under $\mathcal{M}$, then $\mathcal{M}$ transfers all the source data retrieved by $Q_S$. 
Source information transferred by a mapping can be characterized in terms of queries.

**Theorem**

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be specified by FO-to-CQ, then:

$$\mathcal{M}_1 \preceq_s \mathcal{M}_2 \text{ if and only if }$$

every source query that is target rewritable under $\mathcal{M}_1$ is also target rewritable under $\mathcal{M}_2$. 
Source information transferred by a mapping can be characterized in terms of queries.

**Theorem**

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be specified by FO-to-CQ, then:

$$\mathcal{M}_1 \preceq_s \mathcal{M}_2 \text{ if and only if }$$

every source query *that is target rewritable under $\mathcal{M}_1$ is also target rewritable under $\mathcal{M}_2$.*

The characterization is particular for FO-to-CQ. For example, it does not work for CQ-to-UCQ.
Deciding $\leq_S$

**Theorem**

*For mappings specified by FO-to-CQ:*

\[
\text{testing } \mathcal{M}_1 \leq_S \mathcal{M}_2 \text{ is undecidable}
\]
Deciding $\lesssim_S$

<table>
<thead>
<tr>
<th>Theorem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>For mappings specified by FO-to-CQ:</strong></td>
</tr>
<tr>
<td>$\text{testing } \mathcal{M}_1 \lesssim_S \mathcal{M}_2 \text{ is undecidable}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theorem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>For mappings specified by CQ$\neq$-to-CQ</strong></td>
</tr>
<tr>
<td>$\text{testing } \mathcal{M}_1 \lesssim_S \mathcal{M}_2 \text{ is decidable}$</td>
</tr>
</tbody>
</table>
Deciding $\leq_S$

**Theorem**

For mappings specified by FO-to-CQ:

\[
\text{testing } M_1 \leq_S M_2 \text{ is undecidable}
\]

**Theorem**

For mappings specified by CQ$\neq$-to-CQ

\[
\text{testing } M_1 \leq_S M_2 \text{ is decidable}
\]

**Proof idea**

For CQ$\neq$-to-CQ mappings, we prove that:

- checking target rewritability for UCQ$\neq$ is decidable,
- only a finite number of queries in UCQ$\neq$ need to be checked to determine if $M_1 \leq_S M_2$. 
Application: Invertibility can be characterized using $\preceq_s$.

Let $\overline{\text{id}}$ be a mapping specified by a set of *copying* (rules of the form $R(\bar{x}) \rightarrow \hat{R}(\bar{x})$ with $R$ a source relation).
Application: Invertibility can be characterized using $\preceq_s$.

Let $\overline{\text{id}}$ be a mapping specified by a set of *copying* (rules of the form $R(\overline{x}) \rightarrow \hat{R}(\overline{x})$ with $R$ a source relation).

Definition [F06]: $\mathcal{M}'$ is an *inverse* of $\mathcal{M}$ if $\mathcal{M} \circ \mathcal{M}' = \overline{\text{id}}$. 
Application: Invertibility can be characterized using \( \preceq_s \).

Let \( \overline{Id} \) be a mapping specified by a set of *copying* (rules of the form \( R(\overline{x}) \rightarrow \hat{R}(\overline{x}) \) with \( R \) a source relation).

Definition [F06]: \( M' \) is an *inverse* of \( M \) if \( M \circ M' = \overline{Id} \).

**Theorem**

Consider the class of total and closed-down on the left mappings:

- \( M \) is invertible \( \iff \overline{Id} \preceq_s M \)
- \( M \) is invertible \( \iff M \) is \( \preceq_s \)-maximal
Application: Invertibility can be characterized using $\preceq_s$.

Let $\overline{Id}$ be a mapping specified by a set of *copying* (rules of the form $R(\overline{x}) \to \hat{R}(\overline{x})$ with $R$ a source relation).

**Definition [F06]:** $\mathcal{M}'$ is an *inverse* of $\mathcal{M}$ if $\mathcal{M} \circ \mathcal{M}' = \overline{Id}$.

**Theorem**

*Consider the class of total and closed-down on the left mappings:*

- $\mathcal{M}$ is invertible $\iff \overline{Id} \preceq_s \mathcal{M}$
- $\mathcal{M}$ is invertible $\iff \mathcal{M}$ is $\preceq_s$-maximal

Invertibility do coincide with transferring the maximum amount of source information!
Application: Invertibility can be characterized using $\leq_s$.

Let $\overline{Id}$ be a mapping specified by a set of copying (rules of the form $R(\overline{x}) \rightarrow \hat{R}(\overline{x})$ with $R$ a source relation).

Definition [F06]: $M'$ is an inverse of $M$ if $M \circ M' = \overline{Id}$.

**Theorem**

Consider the class of total and closed-down on the left mappings:

- $M$ is invertible $\iff$ $\overline{Id} \leq_s M$
- $M$ is invertible $\iff$ $M$ is $\leq_s$-maximal

Invertibility do coincide with transferring the maximum amount of source information!

**Corollary [FN09]**

Testing invertibility for $\text{CQ} \neq \text{-to-CQ}$ mappings is decidable.
Covering target information: the *dual* definition

Assume that $\mathcal{M}_1$ and $\mathcal{M}_2$ share the target schema.

**Definition**

$\mathcal{M}_2$ is *more (or equally) target-informative* than $\mathcal{M}_1$, denoted by

\[
\mathcal{M}_1 \preceq_T \mathcal{M}_2,
\]

if there exists a mapping $\mathcal{M}'$ such that $\mathcal{M}' \circ \mathcal{M}_2 = \mathcal{M}_1$. 


Covering target information: the dual definition

Assume that $M_1$ and $M_2$ share the target schema.

**Definition**

$M_2$ is more (or equally) target-informative than $M_1$, denoted by $M_1 \preceq_T M_2$, if there exists a mapping $M'$ such that $M' \circ M_2 = M_1$.

► *Universal solutions* [FKMP05]: A solution $J^*$ for an instance $I$ that represents the entire space of solutions of $I$ under $M$. 
Covering target information: the *dual* definition

Assume that $\mathcal{M}_1$ and $\mathcal{M}_2$ share the target schema.

**Definition**

$\mathcal{M}_2$ is *more (or equally) target-informative* than $\mathcal{M}_1$, denoted by $\mathcal{M}_1 \preceq_T \mathcal{M}_2$, if there exists a mapping $\mathcal{M}'$ such that $\mathcal{M}' \circ \mathcal{M}_2 = \mathcal{M}_1$.

▶ *Universal solutions* [FKMP05]: A solution $J^*$ for an instance $I$ that represents the *entire space of solutions* of $I$ under $\mathcal{M}$.

**Theorem**

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be specified by *FO-to-CQ*, then:

$$\mathcal{M}_1 \preceq_T \mathcal{M}_2 \text{ if and only if every target instance that is universal solution under } \mathcal{M}_1 \text{ is also universal solution under } \mathcal{M}_2.$$
Application 2: formalization of Extract (first attempt)

We model the extract of $\mathcal{M}$ as a pair $(\mathcal{M}_1, \mathcal{M}_2)$ s.t.
Application 2: formalization of \textit{Extract} (first attempt)

We model the \textit{extract} of $\mathcal{M}$ as a pair ($\mathcal{M}_1, \mathcal{M}_2$) s.t.

(E1) $\mathcal{M}_1 \equiv_s \mathcal{M}$
We model the extract of $\mathcal{M}$ as a pair $(\mathcal{M}_1, \mathcal{M}_2)$ s.t.

\[(E1) \quad \mathcal{M}_1 \equiv_{S} \mathcal{M} \quad \text{(i.e. } \mathcal{M}_1 \preceq_{S} \mathcal{M} \text{ and } \mathcal{M} \preceq_{S} \mathcal{M}_1)\text{.}\]
We model the *extract* of $\mathcal{M}$ as a pair $(\mathcal{M}_1, \mathcal{M}_2)$ s.t.

(E1) $\mathcal{M}_1 \equiv_s \mathcal{M}$ (i.e. $\mathcal{M}_1 \preceq_S \mathcal{M}$ and $\mathcal{M} \preceq_S \mathcal{M}_1$).

(E2) $\mathcal{M}_2 \equiv_t \mathcal{M}$
Application 2: formalization of Extract (first attempt)

We model the extract of $\mathcal{M}$ as a pair $(\mathcal{M}_1, \mathcal{M}_2)$ s.t.

(E1) $\mathcal{M}_1 \equiv_s \mathcal{M}$ (i.e. $\mathcal{M}_1 \preceq_s \mathcal{M}$ and $\mathcal{M} \preceq_s \mathcal{M}_1$).

(E2) $\mathcal{M}_2 \equiv_t \mathcal{M}$ (i.e. $\mathcal{M}_2 \preceq_t \mathcal{M}$ and $\mathcal{M} \preceq_t \mathcal{M}_2$).
Application 2: formalization of Extract (first attempt)

We model the extract of $\mathcal{M}$ as a pair $(\mathcal{M}_1, \mathcal{M}_2)$ s.t.

(E1) $\mathcal{M}_1 \equiv_s \mathcal{M}$ (i.e. $\mathcal{M}_1 \preceq_s \mathcal{M}$ and $\mathcal{M} \preceq_s \mathcal{M}_1$).
(E2) $\mathcal{M}_2 \equiv_t \mathcal{M}$ (i.e. $\mathcal{M}_2 \preceq_t \mathcal{M}$ and $\mathcal{M} \preceq_t \mathcal{M}_2$).
(E3) $\mathcal{M} = \mathcal{M}_1 \circ \mathcal{M}_2$
Application 2: formalization of \textit{Extract} (first attempt)

We model the \textit{extract} of $\mathcal{M}$ as a pair $(\mathcal{M}_1, \mathcal{M}_2)$ s.t.

\begin{itemize}
  \item [(E1)] $\mathcal{M}_1 \equiv_S \mathcal{M}$ \quad (i.e. $\mathcal{M}_1 \preceq_S \mathcal{M}$ and $\mathcal{M} \preceq_S \mathcal{M}_1$).
  \item [(E2)] $\mathcal{M}_2 \equiv_T \mathcal{M}$ \quad (i.e. $\mathcal{M}_2 \preceq_T \mathcal{M}$ and $\mathcal{M} \preceq_T \mathcal{M}_2$).
  \item [(E3)] $\mathcal{M} = \mathcal{M}_1 \circ \mathcal{M}_2$
\end{itemize}
Application 2: formalization of *Extract* (first attempt)

We model the *extract* of $\mathcal{M}$ as a pair $(\mathcal{M}_1, \mathcal{M}_2)$ s.t.

- (E1) $\mathcal{M}_1 \equiv_s \mathcal{M}$ (i.e. $\mathcal{M}_1 \preceq_s \mathcal{M}$ and $\mathcal{M} \preceq_s \mathcal{M}_1$).
- (E2) $\mathcal{M}_2 \equiv_t \mathcal{M}$ (i.e. $\mathcal{M}_2 \preceq_t \mathcal{M}$ and $\mathcal{M} \preceq_t \mathcal{M}_2$).
- (E3) $\mathcal{M} = \mathcal{M}_1 \circ \mathcal{M}_2$

¿How do we ensure the *optimality* of the new source schema?
Outline

Motivation

Source information
   Algorithmic issues
   Application: Invertibility

Target information
   Application: Extract, first approach

Target and source redundancy
   Application: Extract

Concluding remarks
Target redundancy in mappings: Intuition

Source: \{\text{Emp}(\text{name, lives\_in, works\_in}) \}
Target redundancy in mappings: Intuition

Source: \{\text{Emp}(\text{name}, \text{lives\_in}, \text{works\_in}) \}\n
Target_1: \{\text{ENames}(\text{name}), \text{WorksIn}(\text{name}, \text{place}) \}\
Target redundancy in mappings: Intuition

Source: \{\text{Emp(name, lives\_in, works\_in)} \}
Target\_1: \{\text{ENames(name), WorksIn(name, place)} \}

\[ \mathcal{M}_1: \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z) \]
Target redundancy in mappings: Intuition

Source: \{\text{Emp}(\text{name, lives_in, works_in}) \}\n
Target_1: \{\text{ENames(name), WorksIn(name, place)} \}\n
Target_2: \{\text{Worker(name, working_place)} \}\n
\mathcal{M}_1: \quad \text{Emp}(x, y, z) \quad \rightarrow \quad \text{ENames}(x) \land \text{WorksIn}(x, z)
Target redundancy in mappings: Intuition

Source: \{\text{Emp(name, lives\_in, works\_in)} \}

Target$_1$: \{\text{ENames(name)}, \text{Works\_In(name, place)} \}

Target$_2$: \{\text{Worker(name, working\_place)} \}

\[ M_1: \quad \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{Works\_In}(x, z) \]

\[ M_2: \quad \text{Emp}(x, y, z) \rightarrow \text{Worker}(x, z) \]
Target redundancy in mappings: Intuition

Source: \{Emp(name, lives_in, works_in) \}
Target\textsubscript{1}: \{ENames(name), WorksIn(name, place) \}
Target\textsubscript{2}: \{Worker(name, working_place) \}

\[ M_1: \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z) \]
\[ M_2: \text{Emp}(x, y, z) \rightarrow \text{Worker}(x, z) \]

Intuitively:

\( \bullet \) \( M_1 \) is target redundant:

employee names are stored twice in the target schema.
Target redundancy in mappings: Intuition

Source: \{\text{Emp(name, lives\_in, works\_in)} \}

Target\_1: \{\text{ENames(name)}, \text{WorksIn(name, place)} \}

Target\_2: \{\text{Worker(name, working\_place)} \}

\[ M_1: \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z) \]

\[ M_2: \text{Emp}(x, y, z) \rightarrow \text{Worker}(x, z) \]

Intuitively:

\( M_1 \) is \textit{target redundant}:
employee names are stored twice in the target schema.

\( M_2 \) is \textit{not target redundant}:
all information in the target is \textit{essential} for \( M_2 \).
Target redundancy in mappings: Intuition

Source: \{ \text{Emp(name, lives\_in, works\_in)} \}

Target\(_1\): \{ \text{ENames(name)}, \text{WorksIn(name, place)} \}

Target\(_2\): \{ \text{Worker(name, working\_place)} \}

\[ M_1: \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z) \]

\[ M_2: \text{Emp}(x, y, z) \rightarrow \text{Worker}(x, z) \]

Intuitively:

- \( M_1 \) is target redundant:
  employee names are stored twice in the target schema.

- \( M_2 \) is not target redundant:
  all information in the target is essential for \( M_2 \).

Notice that \( M_1 \) and \( M_2 \) are equally source-informative, \( M_1 \equiv_s M_2 \)
Target redundancy in mappings: Formalization

Definition

\( \mathcal{M} \) is target redundant if there is an instance \( J^* \in \text{range}(\mathcal{M}) \) such that the mapping

\[ \mathcal{M}' = \{(I, J) \in \mathcal{M} \mid J \neq J^*\} \]

and \( \mathcal{M} \) are equally source-informative (\( \mathcal{M} \equiv_s \mathcal{M}' \)).
Target redundancy in mappings: Formalization

**Definition**

\( \mathcal{M} \) is *target redundant* if there is an instance \( J^* \in \text{range}(\mathcal{M}) \) such that the mapping

\[
\mathcal{M}' = \{(I, J) \in \mathcal{M} \mid J \neq J^*\}
\]

and \( \mathcal{M} \) are equally source-informative (\( \mathcal{M} \equiv_s \mathcal{M}' \)).

We can *lose a target instance*, and still be able to transfer the same amount of source information.
Target redundancy in mappings: Formalization

Definition

\( \mathcal{M} \) is \textit{target redundant} if there is an instance \( J^* \in \text{range}(\mathcal{M}) \) such that the mapping

\[
\mathcal{M}' = \{(I, J) \in \mathcal{M} \mid J \neq J^*\}
\]

and \( \mathcal{M} \) are equally source-informative (\( \mathcal{M} \equiv_s \mathcal{M}' \)).

We can \textit{lose a target instance}, and still be able to transfer the same amount of source information.

\[\mathcal{M}_1: \text{Emp}(x, y, z) \rightarrow \text{ENames}(x) \land \text{WorksIn}(x, z)\]

\( J^* \):

<table>
<thead>
<tr>
<th>ENames:</th>
<th>name</th>
<th>place</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cristian</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( J^* \):

<table>
<thead>
<tr>
<th>WorksIn:</th>
<th>name</th>
<th>place</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juan</td>
<td></td>
<td>Santiago</td>
</tr>
</tbody>
</table>
Source redundancy in mappings: the dual definition

Definition

\( \mathcal{M} \) is source redundant if there is an instance \( I^* \in \text{dom}(\mathcal{M}) \) such that the mapping

\[
\mathcal{M}' = \{(I, J) \in \mathcal{M} \mid I \neq I^*\}
\]

and \( \mathcal{M} \) are equally target-informative (\( \mathcal{M} \equiv_T \mathcal{M}' \)).
Source redundancy in mappings: the *dual* definition

**Definition**

\( \mathcal{M} \) is *source redundant* if there is an instance \( I^* \in \text{dom}(\mathcal{M}) \) such that the mapping

\[
\mathcal{M}' = \{ (I, J) \in \mathcal{M} \mid I \neq I^* \}
\]

and \( \mathcal{M} \) are equally target-informative (\( \mathcal{M} \equiv_T \mathcal{M}' \)).

**Theorem**

Let \( \mathcal{M} \) be specified by FO-to-CQ, then:

- \( \mathcal{M} \) is target redundant iff there is a target instance that is not a universal solution under \( \mathcal{M} \) (onto mapping [FN09]).

- \( \mathcal{M} \) is source redundant iff there are two source instances with the same space of solutions under \( \mathcal{M} \) (unique solutions property [F06]).
Application 2: formalization of *Extract*

$(M_1, M_2)$ is an *extract* of $M$ iff:

1. $(E1)$ $M_1 \equiv_s M$
2. $(E2)$ $M_2 \equiv_t M$
3. $(E3)$ $M = M_1 \circ M_2$
Application 2: formalization of \textit{Extract}

\[(M_1, M_2) \text{ is an extract of } M \text{ iff:}\]

\begin{align*}
\text{(E1)} & \quad M_1 \equiv_s M \\
\text{(E2)} & \quad M_2 \equiv_t M \\
\text{(E3)} & \quad M = M_1 \circ M_2 \\
\text{(E4)} & \quad M_1 \text{ is not target redundant} \\
\text{(E5)} & \quad M_2 \text{ is not source redundant}
\end{align*}
Application 2: formalization of Extract

$(M_1, M_2)$ is an extract of $M$ iff:

(E1) $M_1 \equiv_s M$
(E2) $M_2 \equiv_t M$
(E3) $M = M_1 \circ M_2$
(E4) $M_1$ is not target redundant
(E5) $M_2$ is not source redundant

Theorem

For mappings specified by FO-to-CQ an extract always exists.
Application 2: formalization of $\textit{Extract}$

$\langle M_1, M_2 \rangle$ is an \textit{extract} of $M$ iff:

\begin{align*}
\text{(E1) } & M_1 \equiv_s M \\
\text{(E2) } & M_2 \equiv_t M \\
\text{(E3) } & M = M_1 \circ M_2 \\
\text{(E4) } & M_1 \text{ is not target redundant} \\
\text{(E5) } & M_2 \text{ is not source redundant}
\end{align*}

\textbf{Theorem}

\textit{For mappings specified by FO-to-CQ an extract always exists.}

In the paper: an algorithm to compute an extract.
Information and redundancy are fundamental notions for schema mappings

In our work:

- we provide a formalization for both notions
- we study algorithmic issues, and natural characterizations
- we use these notions to re-study some schema mapping operators (schema evolution, extract, merge, inverse).
Foundations of Schema Mapping Management

Marcelo Arenas, Jorge Pérez, Juan L. Reutter, Cristian Riveros

PUC Chile, U. Edinburgh, U. Oxford
Outline

Motivation

Source information
  Algorithmic issues
  Application: Invertibility

Target information
  Application: Extract, first approach

Target and source redundancy
  Application: Extract

Concluding remarks