unibz
Fakultät für Informatik
Facoltà di Scienze e Tecnologie informatiche
Faculty of Computer Science

Diego Calvanese
Free University of Bozen-Bolzano

Wolfgang Fischl **Reinhard Pichler** **Emanuel Sallinger** **Mantas Šimkus**
Vienna University of Technology

TU WIEN
FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Capturing Functional Dependencies and Relational Schemas in RDFS

Motivation

- Enormous growth in Semantic Web data
- Most Semantic Web data originates from relational databases
- Normal forms for relational databases are important
- GOAL: **Is there a Normal Form for Semantic Web data?**

Semantic Web

- Data is stored in RDF graphs extended with RDFS
- Focus on *DL-Lite*$_{\textbf{RDFS}}$ with disjointness

Relational Databases

- Data is stored in n-ary relations
- Functional dependencies (FDs) ensure *consistency*
- Normal forms avoid *redundancy*

Direct Mapping of Relational Data to RDFS Graph Data

- each database instance corresponds to a model of the ontology and vice versa.

Capturing Functional Dependencies

- the translation preveserves satisfaction of dependencies

Normal Form for Semantic Web Data

- Boyce-Codd Normal Form (BCNF) in the Semantic Web with
- tree-based identification constraints.

Example

Table: ***loc***

room	***house***	***addr***
Entrance	White_House	1600_PA_Av
Oval_Office	White_House	1600_PA_Av

1. Translate schema information

- Concept disjointness:

$$\textbf{\textit{locs}} \sqsubseteq \neg \textbf{\textit{room}} \qquad \textbf{\textit{room}} \sqsubseteq \neg \textbf{\textit{house}}$$
$$\textbf{\textit{room}} \sqsubseteq \neg \textbf{\textit{addr}} \qquad \textbf{\textit{house}} \sqsubseteq \neg \textbf{\textit{addr}}$$

- Role typing:

$$\exists \textbf{\textit{room}} \sqsubseteq \textbf{\textit{has\_room}} \qquad \exists \textbf{\textit{has\_room}}^- \sqsubseteq \textbf{\textit{room}}$$
$$\exists \textbf{\textit{house}} \sqsubseteq \textbf{\textit{has\_house}} \qquad \exists \textbf{\textit{has\_house}}^- \sqsubseteq \textbf{\textit{house}}$$
$$\exists \textbf{\textit{addr}} \sqsubseteq \textbf{\textit{has\_addr}} \qquad \exists \textbf{\textit{has\_addr}}^- \sqsubseteq \textbf{\textit{addr}}$$

- Functionality and reification assertions:

$$(\textbf{funct } \textbf{\textit{has\_room}})$$
$$(\textbf{funct } \textbf{\textit{has\_house}})$$
$$(\textbf{funct } \textbf{\textit{has\_addr}})$$
$$\textbf{\textit{loc}}?(\textbf{\textit{has\_room}}, \textbf{\textit{has\_house}}, \textbf{\textit{has\_addr}})$$

2. Translate data

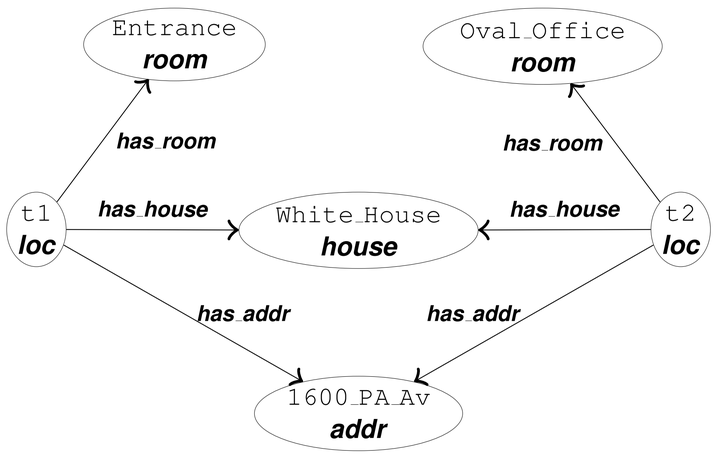


Figure 1: Translated relational table

- Use path-based identification constraints (pids) *[Calvanese et al., KR 2008]
- There are FDs for which the above property does not hold for any set of pids.

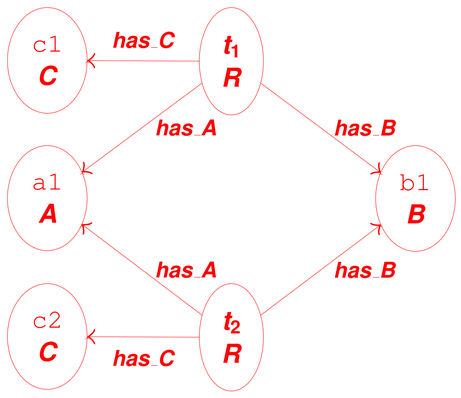
Example

- Consider the FD $\{\textbf{\textit{A}}, \textbf{\textit{B}}\} \rightarrow \textbf{\textit{C}}$ and
- its translation to a pid

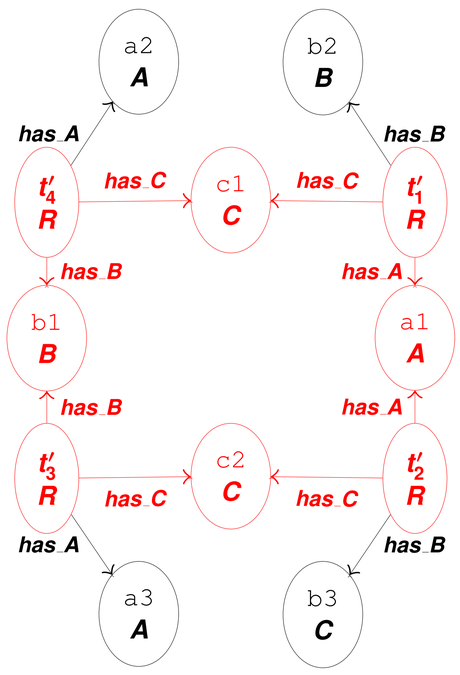
$$\textbf{\textit{C}}?(\textbf{\textit{has\_C}}^- \cdot \textbf{\textit{has\_A}}, \textbf{\textit{has\_C}}^- \cdot \textbf{\textit{has\_B}}).$$

- Translation *does not* preserve satisfaction of dependencies.

	A	***B***	***C***
t_1	a1	b1	c1
t_2	a1	b1	c2



	A	***B***	***C***
t'_1	a1	b2	c1
t'_2	a2	b1	c1
t'_3	a3	b1	c2
t'_4	a4	b3	c3



FD and pid violation in red.

- BCNF avoids update anomalies originating
- from redundancies induced by FDs.

Example

- ***addr***?(***has_addr***$^-$ · ***has_house***)
- leads to a redundancy.
- Avoided by storing ***addr*** directly to the ***house***.

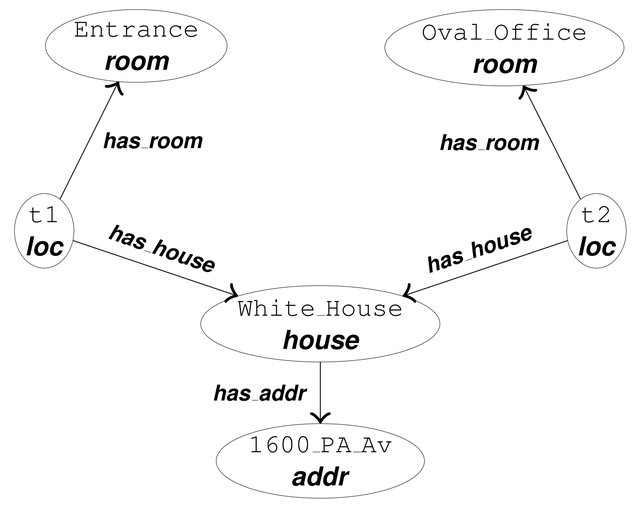


Figure 2: An RDF graph in RNF

Definition

An RDF graph is in *RDF Normal Form* (RNF) if and only if all tids can be substituted by tids with depth 1.

Theorem

A relational database is in BCNF iff the RDF graph with tids (translated with the direct mapping) is in RNF.

Theorem

Deciding whether an RDF graph is in RNF is feasible in polynomial time.

Future Work

- Extend the definition of RNF to more expressive DLs and
- investigate in these DLs the recognizability of RNF.
- Consider relaxations of our RNF definition.
- Investigate the complexity of reasoning with tids.
- Compare RNF with normal forms of other non-relational data formats (such as XML Normal Form).

Tree-based identification constraints (tids)

Definition

A tid is an expression constructed using the following grammar:

$$\tau ::= \epsilon \mid \sigma \cdot \tau \mid (\tau, \tau)$$

where σ is an ordinary role or a test role of the form ***C***?, and ***C*** is a concept.

Example

The tid $\textbf{\textit{C}}?(\textbf{\textit{has\_C}}^- \cdot (\textbf{\textit{has\_A}}, \textbf{\textit{has\_B}}))$ captures the FD $\{\textbf{\textit{A}}, \textbf{\textit{B}}\} \rightarrow \textbf{\textit{C}}$.

Theorem

Tree-based identification constraints are able to express FDs over RDF graphs.

The first author has been supported by the Wolfgang Pauli Institute Vienna, and by the EU IP project Optique(grant agreement n. FP7-318338). The remaining authors have been supported by the Austrian Science Fund (FWF) project P25207-N23 and P25518-N23, and by the Vienna Science and Technology Fund (WWTF) project ICT12-15.

WWTF
FWF Der Wissenschaftsfonds.