MAGIK: Managing Completeness of Data

http://magik-demo.inf.unibz.it/public-version/

Second European Business Intelligence Summer School (eBISS 2012)
July 15 – 21, 2012 Brussels, Belgium

Ognjen Savkovic
Free University of Bozen-Bolzano, Italy
savkovic@inf.unibz.it
Supervisor: **Werner Nutt**

Freie Universität Bozen
Libera Università di Bolzano
Free University of Bozen · Bolzano

Data Quality: State of the art

Data Quality?

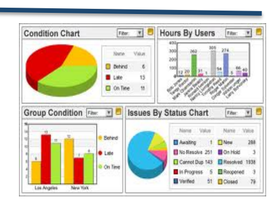
- **Data Quality (DQ) is a perception of data's fitness to serve its purpose in a given context.**
- DQ problem costs U.S. businesses around **$600 billion annually** [TWDI Journal]
- DQ is characterized by (**no silver bullet for DQ problems**) dimensions**: Correctness, Consistency, Accurate, Timeliness …**

Data Completeness?

- **Is all necessary data present?**
- **Measure:** The extent to which data are of sufficient breadth, depth, and scope for the task at hand, f.e. **query answering.**

Data quality in for Decision Support Data

- **Managers** see the data model throughout **Dashboards**.
- **Data quality as a major concern** of decision support data (interviews with IT experts companies and school administration)
- **Diffuse market**: very expensive and non-standardized solutions



Research path

Goal: Take advantage of widely present and formalized methodologies and technologies, that provide **meta-information**, like **Business Processes workflows, Master Data Management**, **etc.** to assess data quality aspects. Track back **the causes of bad data quality and propose solutions. Implement and test** systems that automatize all that.

Motivation Examples

Schema

pupil(name,level,code) … *pupils*
class(level,code,dept) … *every class belongs to a department*
langAtt(name,language) … *pupils attend language courses*

Plain reasoning

TC Statement 1: *We are complete for all pupils.*
`TABLE: pupil(Name,Level,Class) WHERE:`

TC Statement 2: We are complete for all pupils in the class 1a.
`pupil(Name,Level,Class) WHERE: Level=1 AND Class='a'`

Query 1: *Who are the pupils at the 1st class?*

```
SELECT p.name
FROM pupil AS p
WHERE p.level='1'
```

*Can we answer **Query 1** completely under the assumption of **Statement 1**?*

Query 1 is complete, because it so more "specific" then Statement 1.

*Can we answer **Query 1** completely under the assumption of **Statement 2**?*

Query 1 is NOT complete, because it so more "general" then Statement 2.

Reasoning under Foreign keys

FK 1: `pupil(level,code) REFERENCES class(level,code)`
FK 2: `langAtt(name) REFERENCES pupil(name)`

TC Statement 3 *We are complete for French learners.*
`TABLE: langAtt(Name,Lang) WHERE: Lang='french'`

Query 2: *Which science pupils learns French?*

```
SELECT p.name
FROM pupil AS p, class AS c, langAtt AS l
WHERE p.name=l.name AND l.lang='french'
AND p.level=c.level AND p.code=c.code
AND c.branch='science'
```

Can we answer **Query 2** completely under the assumption of **Statement 3** and Foreign keys **FD1** and **FD2**?
Query 1 is complete, because fks guarantee that for every language learner exists a pupil and a class record.

Reasoning under Finite Domains (FD)

FD 1: *Codes of the pupils classes can be either **a** or **b**.*
`pupil(code) IN {a,b}`

TC Statement 2: *We are complete for all pupils in the class **1a**.*
`TABLE: pupil(Name,Level,Class) WHERE: Level=1 AND Class='a'`

Can we answer **Query 1** completely under the assumption of **Statement 2** and **FD1**?
Query 1 is NOT complete, because other classes, like 1b can exists.

What is incomplete wrt **Query 1**? `WHERE: Level=1 AND Class='b'`

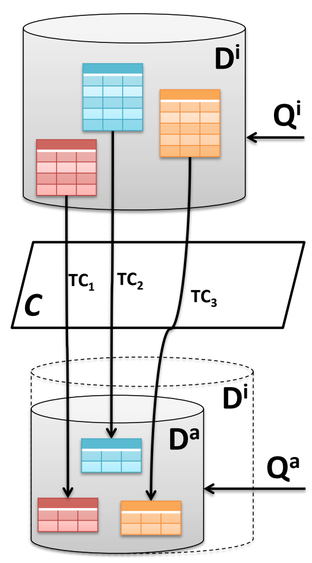
What MAGIK suggests to us ( )? `TABLE:pupil(Name,Level,Class) WHERE:Level=1 AND Class='b'`

*Is **Query 1** complete if we in addition consider the* TC-statement proposed by MAGIK?
Query 1 is complete, because we are complete for all possible 1st classes.

Problem Statement

Given a information (meta-statements , called Table Completeness (TC) statements) that some parts of available database (D^a) is complete can we guarantee (deduce) that a query answer is the same (called Query Completeness (QC)) as the query is evaluated over the complete (ideal) database (D^i)?

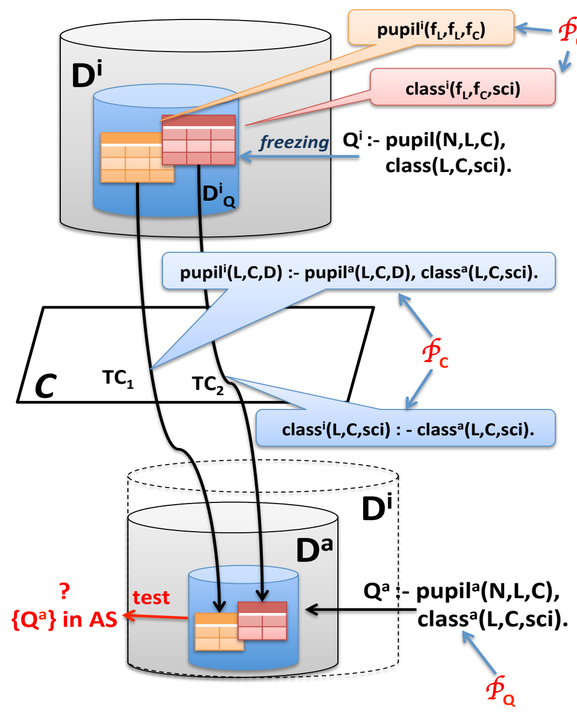
- To express partial completeness of database we use **table (local) completeness (TC)** statements [H.Levy '96]
- tc_1: *We are complete for* **science** *pupils.*
  - `TABLE: pupil(Name,Level,Code) WHERE: class(Level,Code,science).`
  - $\texttt{pupil}^a(N,L,C) \leftarrow \texttt{pupil}^i(N,L,C) \wedge \texttt{class}^i(L,C,\text{science})$.
- Let $D^a = \{\texttt{class(1,a,sci)}\}$, $D_1^i = \{\texttt{class(1,a,sci)}\}$ and $D_2^i = \{\texttt{class(1,a,sci)}, \texttt{pupil(john,1,a)}\}$
  - (D^a, D_1^i) **satisfies** tc_1
    (D^a, D_2^i) **doesn't satisfy** tc_1
  - Similarly for a query $Q(N) \leftarrow \texttt{pupil(N,L,C)}$:
    Q **is complete** under (D^a, D_1^i)
    Q **is NOT complete** under (D^a, D_2^i)

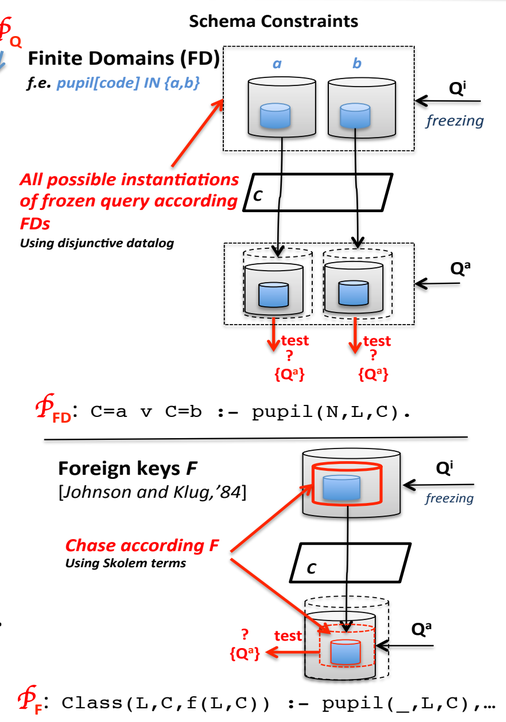


Goal: **Automate the reasoning on query completeness (QC) given information about complete parts of a database (TC-statements).**

Implementation

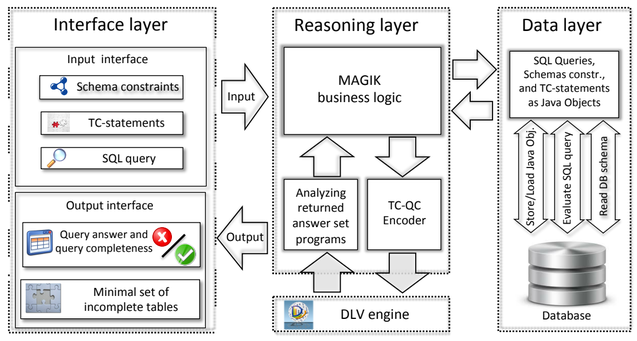
Encoding of the Problem in Logic Programming (Answer Set Programming)





$$\mathcal{P}_C \cup \mathcal{P}_\mathcal{F} \cup \mathcal{P}_{FD} \cup \mathcal{P}_Q \models \{Q^a\} \text{ iff } \mathcal{C} \models_{\mathcal{F},FD} \mathit{Compl}(Q)$$

System Architecture



Summary

- The **first realized** system that can reason about reasons on **query completeness** based on the **partial database (table) completeness (TC-QC)**
- We gone beyond original **TC-QC problem**, and we investigate the impact of **Schema Constraints**, like **Foreign keys** and **Finite Domains**, on TC-QC entailment.
- We developed a component for **explanations and suggestions**, in the case the query is not complete, that indicates which parts of a database are incomplete wrt the query.