



## Topic for Bachelor's / Master's Thesis

In the department of computer science / research group of database and information systems, we offer the following topic for a bachelor's / master's thesis:

### Ranking Constraint Violations in Knowledge Bases

#### Motivation

Knowledge bases such as Wikidata are used for a wide range of applications, e.g., quick answer boxes in search engines (e.g. Google, Bing), personal assistants (e.g., Siri, Google), or question answering systems (e.g., IBM Watson). However, today's knowledge bases suffer from quality problems. For example, Wikidata currently reports over 10 million constraint violations. In traditional databases, all data which violates constraints is simply discarded. However, this approach is not applicable to real-world, large-scale knowledge bases as for almost every constraint, there is an exception in the real world, and strictly enforcing constraints prevents an agile and flexible development of the knowledge base. Nevertheless, constraint violations often point to quality problems. To overcome this dilemma, we envision a semi-automatic approach: constraint violations are ranked by the severity of their consequences, thus, enabling the volunteers of the knowledge base to manually review and fix the most important violations first.

#### Description of the Task

- Investigate some examples of constraint violations in Wikidata, and manually order them by the severity of their consequences
- Develop systematic criteria to rank constraint violations in knowledge bases (more/better criteria for master's thesis)
- Develop a prototype for automatically ranking the constraint violations
- Evaluate your prototype by comparing its result with your initial, manual ranking (or even perform a crowdsourcing experiment for master's thesis)
- For the most common and severe types of constraint violations, offer suggestions how to fix them (semi-) automatically

#### Result for Barack Obama (Q76)

Violation: 22, Compliance: 636, Todo: 153

Status	Claim	Constraint
Violation [?]	ethnic group: Kenyan American	Single value
Violation [?]	Twitter username: BarackObama	Single value
Violation [?]	official website: <a href="http://www.whitehouse.gov/administration/president_obama/">http://www.whitehouse.gov/administration/president_obama/</a>	Single value
Violation [?]	PTBNP ID: 1378719	Single value
Violation [?]	part of: 109th United States Congress	Inverse [...]
Violation [?]	official website: <a href="http://www.barackobama.com/">http://www.barackobama.com/</a>	Single value
Violation [?]	medical condition: acid reflux disease	Value type [...]
Violation [?]	part of: 110th United States Congress	Inverse [...]
Violation [?]	PTBNP ID: 1470399	Single value
Violation [?]	part of: Congressional Black Caucus	Inverse [...]

<https://www.wikidata.org/wiki/Special:ConstraintReport/Q76>

#### Contact

Stefan Heindorf  
E-Mail: [heindorf@uni-paderborn.de](mailto:heindorf@uni-paderborn.de)  
Office: ZM1.03-07  
Phone: (+49) (0)5251 5465-207