



## 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:

### The History and Future of Data Quality in Knowledge Bases

#### Motivation

Knowledge bases such as DBpedia, Wikidata, or Yago 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, the error-rate is estimated to be between 5 % and 10 %, and the data is often incomplete. Over the last couple of years all major knowledge bases took countermeasure to improve data quality. For example, DBpedia developed better extractors to extract information from Wikipedia, Yago combined the data from many Wikipedias, Wikidata introduced semi-automatic editing tools. However, it has not been systematically studied how all those measures affected data quality and what we can learn from this for the future.

#### Description of the Task

- Identify major events in the history of DBpedia, Wikidata, and Yago which potentially affected data quality (only for Wikidata if Bachelor's thesis)
- Identify important quality metrics that might have been affected
- Develop a prototype to compute quality metrics in knowledge bases over time
- Interpretate your results and compile 'lessons learned' to improve the data quality of knowledge bases in the future

#### Prerequisites

- Interest in big data and scalable systems

#### Contact

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

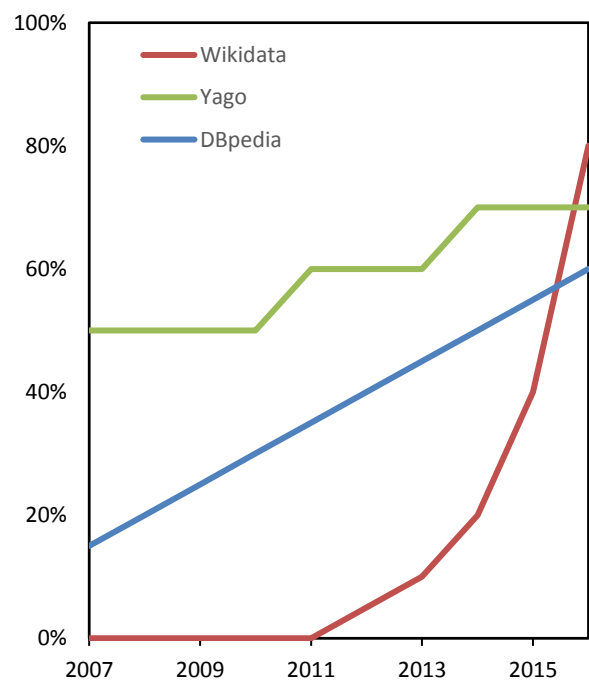


Figure: Data Quality in Knowledge Bases over Time (fictitious values)