The Quest to Know What We Know

PI: Simon Razniewski
Co-PI: Werner Nutt
External partner: Fabian Suchanek, Telecom ParisTech

October 14, 2015

Abstract

Recently arising knowledge bases collect a vast number of facts about the world. But while thus quite some facts are known about the world, little is known about how much is actually known. In this project we aim to lay the foundations for estimating how much knowledge bases know.

We intend to proceed threefold: First, we want to create a labelled gold-standard data set, where we use crowdsourcing to label for a small topic how much a knowledge base knows about this topic. Second, we would experiment with rule-mining techniques to see how they can be used to infer completeness of entities based on completeness information about similar entities. Third, we intend to develop presentation techniques for information about how much is known about the world.

1 Motivation

General-purpose knowledge bases (KBs) such as Wikidata [19], the Google Knowledge Vault [3], NELL [10] or YAGO [18] aim to collect as much factual information about the world as possible. Both how much they know and the correctness of the facts they know, called recall and precision, are important aspects of these KBs, however, commonly more emphasis is placed on precision than on recall (e.g. YAGO cites a goal of 95% precision). While still they aim to provide a high recall, not much is known about their recall.

In fact, that such knowledge bases will ever achieve a high recall may seem an utterly impossible task:

1. DBpedia contains currently only 0.03% of all living persons.
2. According to YAGO, the average number of children per person is 0.02.
3. The Google Knowledge Graph contains for countries a vague predicate called "Points of Interest".
4. In Biperpedia, 67k different attributes can be expressed [7], for which each recall is a question.

Nonetheless, KBs do know very well about certain popular topics:

	#entities	#facts
KB 1	4 mill	72 mill
KB 2	2.3 mill	21 mill
KB 3	17 mill	43 mill

Table 1: State of the art for comparing the recall of KBs

1. 160 out of 199 Nobel laureates in Physics are in DBpedia.
2. 2 out of 2 children of Obama are in Wikidata.
3. 36 out of 48 movies by Tarantino are shown in the Google Knowledge Graph.

Absence of information about the recall of KBs is an issue for several reasons:

1. Philosophical: We do not know what we actually know, and what we don't.
2. Data collection: KB contributors and engineers might not know where to focus their effort. If they know that 39 Nobel laureates in Physics are missing, they could trace and add the missing ones.
3. KB debugging: One does not know when too much data is added. If there is reason to believe that Obama has two children, but a KB contains three, this could be highlighted.
4. Rule learning: KBs are often used for rule induction in order to learn new patterns and facts about the real world. But in order to evaluate learned rules, negative information is needed, which is usually not contained in KBs, but could be inferred from completeness information.
5. Data consumption: Consumers do not know whether a query really retrieves all answers. Also, results of aggregate queries (such as the average number of children per person) and queries with negation cannot be trusted.

State-of-the-art evaluations of the recall of KBs largely only count the number of objects and facts contained in the KB, similarly as shown in Table 1.

In this project, we aim to lay the foundations for fine-grained estimates of the recall of KBs.

2 Describing the Recall of Knowledge Bases

Commonly, KBs are interpreted as follows: The facts that the KB contains are correct, while the facts that are not contained in the KB are unknown, and may or may not be true. This interpretation is called the *open-world assumption*.

Under this interpretation, it is not possible to express that a knowledge base knows everything (is complete) about a topic. Also, while there exists theoretical work about negative information (facts that do not hold) in knowledge bases, none of the state-of-the-art KBs contains negative information.

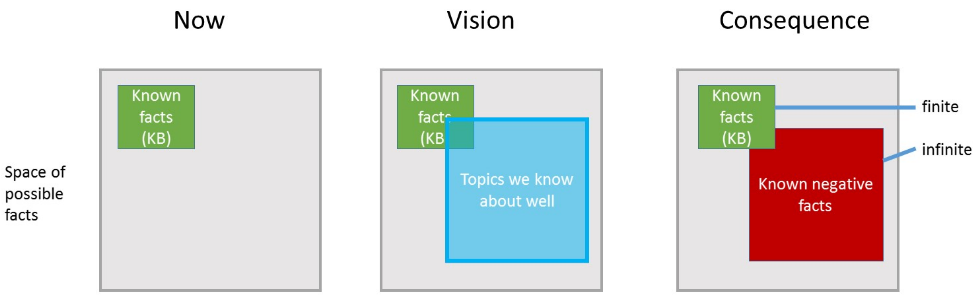


Figure 1: What knowledge about known topics adds technically.

Completeness and negative information are closely related: If we find in a KB that Natasha and Malia are children of Obama, and that the KB contains all children of Obama, we can deduce that anyone else is not a child of Obama, and hence an infinite number of negative facts (see Fig. 1).

Completeness information is also crucial for correct answers to nonmonotonic queries, e.g. queries that include aggregation or negation.

Class Descriptions

Technically, we want to enrich KBs with completeness assertions, which allow to understand how well a KB knows about a certain topic.

An example assertion would be:

This KB knows all children of Obama.

If we had knowledge of this kind, it could be presented alongside the data (Fig. 2), or could be used to annotate query answers (Fig. 3).

Completeness assertions could in principle be given for any definable class of objects. We believe that in particular, classes definable by conjunctive queries are of interest. We call such defining queries, that always have one selection variable *class expressions*.

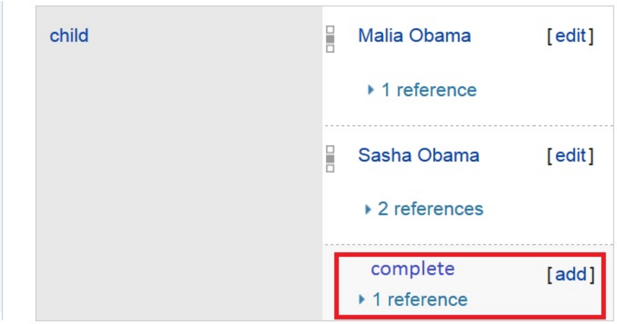
The class expression for children of Obama would be:

```
SELECT c WHERE hasChild(Obama,c).
```

Probabilistic and Quantitative Aspects

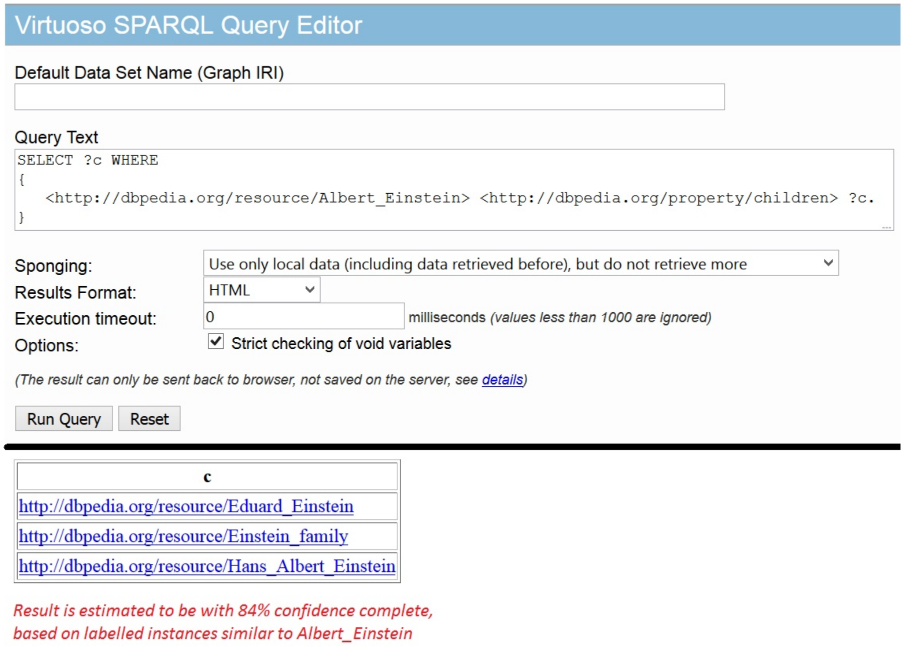
The reliability of completeness information itself may be an issue. Similarly to facts in a KB, it therefore makes sense that completeness information is annotated with *confidences* ("We are 95% sure that the KB contains all children of Obama"), especially if the recall information is automatically extracted.

For many topics, a full *coverage* is difficult or unrealistic. We therefore imagine that completeness can also be expressed as fractions of the entities



child	
	Malia Obama [edit] 1 reference
	Sasha Obama [edit] 2 references
	complete [add] 1 reference

Figure 2: How Wikidata could be annotated with recall (red).



Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

```
SELECT ?c WHERE
{
    <http://dbpedia.org/resource/Albert_Einstein> <http://dbpedia.org/property/children> ?c.
}
```

Sponging: Use only local data (including data retrieved before), but do not retrieve more

Results Format: HTML

Execution timeout: 0 milliseconds *(values less than 1000 are ignored)*

Options: ☑ Strict checking of void variables

(The result can only be sent back to browser, not saved on the server, see details)

c
http://dbpedia.org/resource/Eduard_Einstein
http://dbpedia.org/resource/Einstein_family
http://dbpedia.org/resource/Hans_Albert_Einstein

Result is estimated to be with 84% confidence complete, based on labelled instances similar to Albert_Einstein

Figure 3: How SPARQL query results could be annotated with recall (red).

covered. An example is the statement that DBpedia contains 0.03% of all living persons. We actually computed this coverage based on the known *cardinality* of this class (~7 billion), which in many cases may be easier to obtain.

Confidence and coverage are inversely related: We could at the same time be 80% certain to have at least 30% of the children of John, 40% certain to have at least 50% of his children, and 10% certain to have all his children.

3 State of the Art

Unambiguous Definitions A prerequisite for completeness assertions are unambiguous definitions. Some relations such as "sibling" or "place of birth" are well-defined, while others, such as "hobby" or "affiliation" are not. If a class is not well-defined, completeness has little meaning as well.

In annotation projects in NLP, coannotator agreement is often used as a measure for how unambiguous annotation guidelines are. For classes extracted from the web, we are not aware of any way to guarantee well-definedness.

Retrieving Completeness Information Erxleben et al. [5] have introduced no-values into Wikidata, which are completeness assertions in the special case that a property has no values (e.g. Elisabeth I has no children). Thus, they have introduced a way to crowdsource the creation of completeness information.

In rule learning, the *partial completeness assumption* has been proven to do well in providing negative information [6, 3]. It assumes that whenever a KB contains one pair of property and object for a given subject, then the KB contains all objects for that given subject and predicate. Consequently, any other objects do not satisfy the given predicate for the given subject. For instance, if a KB contains the fact that Natasha is a child of Obama, then it is assumed that the KB contains all children of Obama, and hence, anyone not stated to be a child of Obama is not.

An evaluation of the validity of the PCA for various predicates is contained in [6], showing that in fact it holds for the "hasChild"-relation only in 27% of all cases, while for "worksAt" it holds in 90% of all cases.

Rule mining itself allows to infer more facts from given facts, and is successfully applied in web-scale knowledge bases [6, 3].

Pattern matching is successfully employed in the extraction of facts from the web [11]. The basic idea is a semisupervised approach to learn phrases that signify facts, such as "x is a child of y" for the fact *hasChild(y,x)* based on some seeds.

Mark and recapture techniques, stemming originally from the ecology domain, have been used in the estimation of cardinalities of search results [17]. The idea is to estimate the size of an unknown set based on the overlap between several samples. Advanced mathematical models exist that also allow to take into account nonindependent samples, as it would likely be the case for most information extracted from the web.

Combining Completeness Information So far there exist techniques to use boolean completeness information to annotate query answers with completeness information [9, 15, 8], but no techniques to deal with quantitative or probabilistic information.

In turn, probabilistic databases are a well-established research field [2], it might be possible to transfer probabilistic reasoning about query completeness into these frameworks.

Regarding the estimation of the recall of KBs, the standard approach today is to count entities and facts (see Table 1 for an illustration).

4 Work Plan

The whole project shall have a duration of 18 months. In order to lay the groundwork for estimating the recall of KBs, we aim to work on the following three packages:

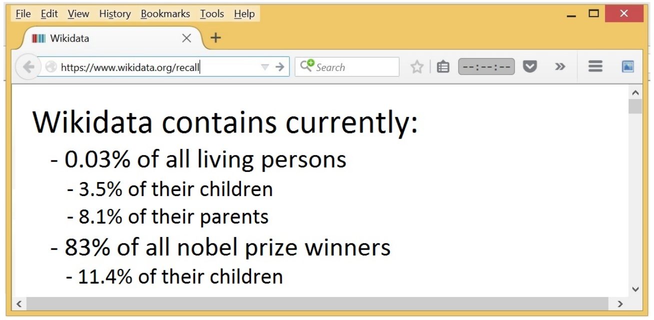
Package 1: Creation of a Gold Standard (Month 1 to 6). In this package, we want to create a labelled gold standard of completeness annotations. We would use a crowdsourcing platform such as the Amazon Mechanical Turk, and label some entities in comprehensible classes such as Nobel prize winners or US presidents the with completeness of some well-defined attributes such as spouses, children, parents. The results of this package would allow a first insight into the recall of existing KBs, and would enable work on the other two packages.

Package 2: Rule-Mining (Month 7 to 18). In this package, we want to apply rule-mining techniques to our labelled data set in order to see how well we can reason about the completeness of unknown instances based on their similarity to instances, where the completeness is known.

Hypotheses we would like to test are (1) whether completeness for different topics of the same entity is correlated (If siblings of a person are complete, are children complete with a higher likelihood too?), and (2) whether completeness across similar entities is correlated (If we know that the children of Obama are complete, are the children of G. W. Bush then complete with a higher likelihood too?).

This package shall be executed in collaboration with Fabian Suchanek from Telecom ParisTech, who is an expert on rule-mining techniques.

Package 3: Presentation of Recall Information (Month 7 to 18). In this package we want to develop a presentation for our completeness annotations, conveying information similarly to the one shown in Fig. 4. The goal would be a comprehensive overview of the completeness of the information from our labelled data set, which will require a way to structure the annotations, provide


File Edit View History Bookmarks Tools Help
Wikidata
https://www.wikidata.org/recall
Search

Wikidata contains currently:
- 0.03% of all living persons
  - 3.5% of their children
  - 8.1% of their parents
- 83% of all nobel prize winners
  - 11.4% of their children

Figure 4: A glimpse of the ultimate goal (in text form).

explanations for shown numbers, and allows zooming in and out wrt. the level of detail.

5 Timeliness and Impact

The field of epistemology, the study of knowledge and beliefs, existed already in the ancient Greece. In turn, the rise of general-purpose KBs is a recent development. For instance, one of the most prominent instances, Wikidata, emerged only in 2012.

While many epistemic questions are open in philosophy, general-purpose KBs bring us close as never before to a comprehensive picture of factual information about the world (remember also IBM's Watson winning over humans at the Jeopardy game, though Watson's knowledge mainly comes from texts, not from KBs).

The publication of a labelled data set with completeness information could be a founding point for wider attention to the problem of recall estimation. Besides, we believe that a study of the extend of factual knowledge in publicly available KBs has potential for popular impact outside of the domain of data management research.

6 Team

Simon Razniewski is an assistant professor at the Free University of Bozen-Bolzano. In his PhD research, for which he has received the best PhD student award in 2013, he has investigated the reasoning about completeness information in various context such as relational databases, business processes, spatial databases and on the semantic web [14]. His work has been published at VLDB [13], CIKM [12], BPM, SIGSPATIAL and SIGMOD [15], and he maintains industrial collaborations with AT&T Labs-Research and Google Research.

Item	Amount (in €)
Assistant for gold-standard creation (3 months)	6000
Assistant for visualization development (3 months)	6000
Data annotation	4000
Collaboration costs	1500
Dissemination costs	2500
Sum	20000

Table 2: Proposed budget.

Fabian M. Suchanek is an associate professor at the Telecom ParisTech University in Paris. Fabian developed inter alia the YAGO-Ontology, which earned him a honorable mention of the SIGMOD dissertation award. His interests include information extraction, automated reasoning, and knowledge bases. Fabian has published around 40 scientific articles, among others at ISWC, VLDB, SIGMOD, WWW, CIKM, ICDE, and SIGIR, and his work has been cited more than 3000 times [18, 6].

Werner Nutt is a full professor at the Free University of Bozen-Bolzano, working on formal aspects of data quality [16] and process governance [16]. In the past, he has been prominently involved in the field of description logic knowledge bases [4, 1].

7 Budget

We intent to hire one assistant (Cocopro) for three months for assisting in the gold-standard creation (6000 €) and one assistant for three month for the development of the visualization (6000 €).

Furthermore, we plan to spend 4000€ for the data annotation, which allows three annotators each to work for one month (8€ per hour), which would allow us to get three labels for around 1000 data items at a rate of 10 minutes per item.

Finally, we intend to use 1500€ for collaboration expenses, e.g. to invite or visit our collaboration partner Fabian Suchanek, and to present our work at a major database conference such as VLDB or SIGMOD (2500€).

The proposed budget is shown in Table 7.

References

[1] Franz Baader and Werner Nutt. Basic description logics. In *Description logic handbook*, pages 43–95, 2003.

[2] Nilesh Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. *The VLDB Journal*, 16(4):523–544, 2007.

[3] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge

vault: a web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 601–610, 2014.

[4] Francesco M Donini, Maurizio Lenzerini, Daniele Nardi, and Werner Nutt. The complexity of concept languages. *Information and Computation*, 134(1):1–58, 1997.

[5] Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandecic. Introducing wikidata to the linked data web. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 50–65, 2014.

[6] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian Suchanek. Fast Rule Mining in Ontological Knowledge Bases with AMIE+. *VLDB Journal*, November 2015.

[7] Rahul Gupta, Alon Y. Halevy, Xuezhi Wang, Steven Euijong Whang, and Fei Wu. Biperpedia: An ontology for search applications. *PVLDB*, 7(7):505–516, 2014.

[8] Willis Lang, Rimma V Nehme, Eric Robinson, and Jeffrey F Naughton. Partial results in database systems. In *SIGMOD*, pages 1275–1286. ACM, 2014.

[9] Alon Y. Levy. Obtaining complete answers from incomplete databases. In *VLDB*, pages 402–412, 1996.

[10] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015.

[11] Dana Movshovitz-Attias, Steven Euijong Whang, Natalya Fridman Noy, and Alon Y. Halevy. Discovering subsumption relationships for web-based ontologies. In *Proceedings of the 18th International Workshop on Web and Databases, Melbourne, VIC, Australia, May 31, 2015*, pages 62–69, 2015.

[12] Werner Nutt and Simon Razniewski. Completeness of queries over SQL databases. In *CIKM*, pages 902–911, 2012.

[13] S. Razniewski and W. Nutt. Completeness of queries over incomplete databases. In *VLDB*, 2011.

[14] Simon Razniewski. Query-driven data completeness management. *CoRR*, abs/1411.2855, 2014.

[15] Simon Razniewski, Flip Korn, Werner Nutt, and Divesh Srivastava. Identifying the extent of completeness of query answers over partially complete databases. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 561–576, 2015.

[16] Ognjen Savkovic, Paramita Mirza, Alex Tomasi, and Werner Nutt. Complete approximations of incomplete queries. *PVLDB*, 6(12):1378–1381, 2013.

[17] Pat Spoor, Mark Airey, Cathy Bennett, Julie Greensill, and Rhys Williams. Use of the capture-recapture technique to evaluate the completeness of systematic literature searches. *BMJ*, 313(7053):342–343, 1996.

[18] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007.

[19] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.