But What Do We Actually Know?

Simon Razniewski
Free University of Bozen-Bolzano
Italy
razniewski@inf.unibz.it

Fabian M. Suchanek
Télécom ParisTech
France
suchanek@enst.fr

Werner Nutt
Free University of Bozen-Bolzano
Italy
nutt@inf.unibz.it

Abstract

Knowledge bases such as Wikidata, DBpedia, YAGO, or the Google Knowledge Vault collect a vast number of facts about the world. But while quite some facts are known about the world, little is known about how much is unknown. For example, while the knowledge base may tell us that Barack Obama is the father of Malia Obama and Sasha Obama, it does not tell us whether these are all of his children. This is not just an epistemic challenge, but also a practical problem for data producers and consumers. We envision that KBs become annotated with information about their recall on specific topics. We show what such annotations could look like, how they could be obtained, and survey related work.

1 Motivation

General-purpose knowledge bases (KBs) such as Wikidata [21], the Google Knowledge Vault [4], NELL [10], or YAGO [19] aim to collect as much factual information about the world as possible. They store information about entities (such as Barack Obama, Hawaii, or NAACL), and information about relationships between these entities (such as the fact that Barack Obama was born in Hawaii, and that NAACL took place in San Diego). These pieces of information typically take the form of triples, as in ⟨*Barack Obama, wasBornIn, Hawaii*⟩. KBs find applications in question answering, automated translation, or information retrieval.

The quality of a KB can be measured along several dimensions. A prominent one is the size. Today's KBs can contain millions, if not billions of triples. Another criterion is precision, i.e., the proportion of triples that are correct. YAGO, e.g., was manually evaluated on a sample, and was shown to have a precision of 95%. In this paper, we propose to look at a third criterion for quality which, besides in some manual evaluations that have ground truth available, has been largely neglected so far: recall, i.e., the proportion of facts of the real world that are covered by the KB. For some topics, today's KBs show very good recall values. For example,

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

On some other topics, today's KBs are nearly completely incomplete:

- DBpedia contains currently only 6 out of 35 Dijkstra Prize winners.
- According to YAGO, the average number of children per person is 0.02.
- The Google Knowledge Graph contains a predicate called "Points of Interest" for countries. Since this predicate is subjective, it is not even clear how to measure its recall.

Previous research [18, 9] has shown that between 69% and 99% of instances in popular KBs lack at least one property that other entities in the same class have. This gives us a hint of how incomplete KBs really are.

The problem is not just that KBs do not contain missing triples, but also that they do not know how

many are missing, or whether some are missing at all. This is an issue from several perspectives:

- Philosophical perspective: We do not know what we actually know, and what we don't.
- Data collection perspective: KB contributors and engineers do not know where to focus their effort. If they knew that 39 Nobel laureates in Physics are missing, they could focus on tracing and adding the missing ones.
- KB debugging perspective: 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.
- Rule learning perspective: 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. Distant supervision, a popular pattern-based technique for automated knowledge base construction, faces the same challenge [9, 15, 20].
- Data consumption perspective: 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.

In this paper, we investigate the problem of recall for KBs, and outline possible approaches to solve it.

2 Vision

Vision. Our vision is that a KB should know for which topics it is complete, and for which topics it is not. Under appropriate interpretation of terms, this could be phrased as

KBs should know what they know.

Defining Completeness. In line with work in databases [11, 8, 13], we define completeness by help of a hypothetical *ideal KB* $\mathcal{K}^*$. The ideal KB contains all facts of the real world. We say that a KB $\mathcal{K}$ is *correct*, if $\mathcal{K} \subseteq \mathcal{K}^*$. We say that $\mathcal{K}$ is *complete* for a query Q, if $Q(\mathcal{K}) \supseteq Q(\mathcal{K}^*)$. For example, we could say that a KB $\mathcal{K}$ is complete for the children of Obama by saying

$\mathcal{K}$ is complete for
`SELECT ?x WHERE {Obama hasChild ?x.}`

This means that evaluating this query on $\mathcal{K}$ will return at least the two children that we would expect as an answer in the real world. Completeness is always bound to a particular query, because we do not expect that we can ever construct a KB $\mathcal{K} = \mathcal{K}^*$. A query can represent the completeness of simple triples about a subject (as in the example), but also for complex constellations, such as "This KB is complete for all rivers longer than 100km in Europe". We believe that completeness assertions are particularly interesting for *class expressions*. These are conjunctive queries with a single selection variable. The class expression for the long rivers of Europe would be:

```
SELECT ?r
WHERE { ?r type river .
        ?r hasLength ?l .
        ?l > 100 .
        ?r locatedIn Europe .}
```

The notion of completeness is closely linked to a number of other concepts, which we detail next.

Closed World Assumption. The *closed world assumption* (CWA) says that if a fact is not in the KB, then it does not hold in the real world. Typically, one restricts this assumption to a certain topic or domain (say, all US presidents). Under the CWA, the KB is always complete for all queries in the domain.

Open World Assumption. Commonly, KBs are not interpreted under the CWA, but under the *open-world assumption* (OWA): The facts that are not in the KB are unknown, and may or may not be true. Under the OWA, we cannot tell whether a KB is complete or not for a given query (unless we have access to $\mathcal{K}^*$).

Negative Information. Negative information (facts that do not hold) is crucial for the correctness of queries with aggregation or negation. While there exists theoretical work about negative information in knowledge bases, none of the state-of-the-art KBs contains negative information. Completeness and negative information are closely related: If we find in a KB that Sasha 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. We thus know an infinite number of negative facts.

Recall. The *recall* of a KB $\mathcal{K}$ for a query Q is $|Q(\mathcal{K}) \cap Q(\mathcal{K}^*)| \times |Q(\mathcal{K}^*)|^{-1}$. The recall is 1 for a

query, if the KB is complete for that query.

Cardinality. The *cardinality* of a query on a KB is the number of results. If we know the cardinality of a query on $\mathcal{K}^*$, and if we know that the KB is correct, we can compute the recall of the KB for that query, and vice versa.

Size. The larger a KB is, the more likely it is to be complete, everything else being equal.

Confidence. Completeness assertions can be crisp, but they could also be made with a certain confidence score. For example, we could be 80% certain to have all children of Obama.

3 Challenges

We see four main challenges that need to be mastered in order to arrive at knowledge about the knowledge of KBs:

3.1 Knowing What Can Be Known

A prerequisite for completeness assertions are unambiguous definitions. Some relations such as "sibling" or "place of birth" are well-defined, while others, such as "affiliation" or "hobby" are not. For example, while one of Einstein's hobbies was playing the violin, he might have had an unclear number of other "hobbies" (such as going for a walk, or eating chocolate). If a topic is not well-defined, completeness has little meaning as well. One might assume that KBs generally contain well-defined predicates, yet this is not always the case. As mentioned before, the Google Knowledge Graph contains an attribute *pointOfInterest*. While some attractions are clearly points of interest (such as the Colosseum in Rome), others are less clearly so (e.g. the pub that DiCaprio allegedly threw up at). In such cases, the concept of crisp completeness is meaningless. We note that some fuzzy concepts can be turned into crisp ones by binding them to particular verifiable properties. For example, it makes sense to consider completeness for "Points of interest recommended by Tripadvisor", because this is a well-defined verifiable set.

3.2 Languages for Describing Completeness

Various formal languages for completeness assertions have been proposed [11, 8, 13], while Erxleben et al. [5] have introduced no-values into Wikidata (e.g. Elisabeth I has no children), thus allowing specifying completeness if the object has no values, but not in the general case. All proposals so far deal only with boolean descriptions, mentioning whether data of some kind is present or not, but do not allow descriptions of confidences or recall.

3.3 Obtaining Completeness Information

Experts. There are two main paradigms for constructing KBs: manual construction by experts or the crowd, and automated extraction from Web sources. For *expert-created data*, it makes sense to give the task of recall estimation to the experts too (as is the case already for the no-values in Wikidata, and wider envisioned in the tool COOL-WD [3]). In this way, a comparable quality of data and recall information can be guaranteed. For *automatically extracted data*, it is highly desirable to find automatic ways to estimate the recall.

Partial Completeness Assumption. The *partial completeness assumption* (PCA) [7] has been proven to do well in providing negative information [7, 4]. It assumes that if a KB contains one pair of property and object for a given subject, then the KB contains all objects for that given subject and property. For instance, if a KB contains the fact that Sasha is a child of Obama, then it is assumed that the KB contains all children of Obama. Hence, anyone who is not known to be a child of Obama is not. The validity of the PCA has been evaluated manually [6] on YAGO. For relations with generally high functionality [17], the PCA holds nearly perfectly. For example, the PCA holds for 90% of the subjects of the *worksAt* relation. For others, the PCA is less suited. For *hasChild*, e.g., the precision of the PCA is only around 27%.

Pattern Matching. Phrases on the Web such as "has no children" or "X and Y are all his children" can be used to infer completeness. Similarly, phrases such as "The 199 Nobel laureates in Physics..." could be used to assert the cardinality and hence the recall for a class.

Growth Patterns over Time. The growth of data over time, and especially the end of such a growth, might indicate completeness. For instance, we can imagine that once a new congress is established, its members are added to a KB until eventually all are inside. The fact that the number then remains constant could indicate completeness.

Interrelation. The completeness of a certain class

expression could be learned from the completeness of other class expressions. For instance, it might be that if parents of a person are complete, then also the children are complete with a higher probability.

Popularity of Entities. It might be that completeness correlates with the number of facts about an entity. For example, if a personality has numerous and very detailed facts in YAGO, then it could be more likely that some basic facts such as his children are complete.

Class Membership. If an entity is a member of a class, we can compare the entity to other members of the class. If other members have attributes that the entity does not have, this could indicate incompleteness. If all class members have the same attributes and the same number of objects, then this could indicate completeness. For example, if all world championships have 32 participants, then also a current championship with 32 participants has a higher probability of being complete.

Similar Entities. If some of the entities are labeled as complete, we could estimate the completeness of other, similar entities. For example, if most football clubs have between 20 and 30 players, and they have been labeled as complete, then a club with 28 players likely has a good recall.

Crowd Sourcing. The crowd could be used to manually generate completeness annotations. A related idea is to use games with a purpose [1].

Estimating Cardinalities. *Mark and recapture* techniques have been developed in the domain of ecology in order to estimate the size of a population of animals. For this purpose, a sample of animals is captured, marked, and freed. After some time, another sample of animals is captured. The ratio of marked animals in this sample can help estimate the size of the population. This technique works also if samples are not independent, and has been used in the estimation of cardinalities of search results [16]. We believe that it might also be useful for estimating the size of a set of entities, based on the overlap between different websites or datasources dealing with the same topic. Based on the size estimate, and the number of entities already in the KB, one could then estimate the recall.

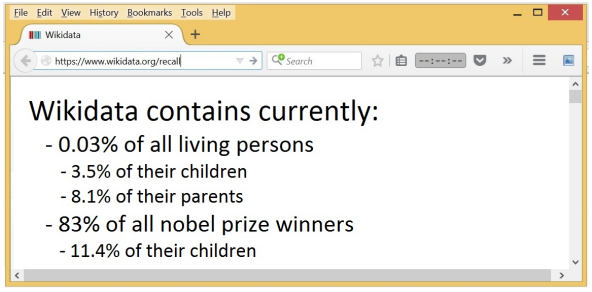

Figure 1: The ultimate vision.

3.4 Combining Completeness Information

Once information about the recall of KBs for individual classes exists, methods need to be found to present this information in a meaningful way. Several techniques can annotate query answers with completeness information [8, 13, 14], but only if the underlying database is annotated with such information, and only for crisp boolean completeness information. Techniques from the domain of query answering over probabilistic databases [2] could possibly be extended to handle non-crisp completeness assertions, while techniques from data profiling can help understand distributions and skew [12].

Also, one would need to apply these techniques to state-of-the-art KBs in order to finally know how much we currently know about the world (see Fig. 1). The community would then have to develop benchmarks for comparing the performance of completeness estimators, and for the completeness of KBs themselves, and would face the classic challenge of KB alignment, because information may be differently presented in different KBs.

4 Conclusion

In this paper, we have outlined our vision of knowledge bases (KBs) that know how complete they are. Their completeness assertions could be used to guide knowledge engineers in the extension and debugging of the KB, to provide negative examples for machine learning algorithms, and to qualify answers to user queries. We have surveyed the state of the art in the area, and concluded that we cannot yet automatically determine where KBs are complete. We have discussed the challenges in defining, determining, and combining completeness assertions, and have outlined possible paths to address them.

Acknowledgement

This work has been partially supported by the projects "TQTK - The Quest to Know", funded by the Free University of Bozen-Bolzano, and "MAGIC", funded by the Province of Bozen-Bolzano.

References

[1] L. v. Ahn. Games with a purpose. *Computer*, 39(6):92–94, June 2006.

[2] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *VLDB Journal*, 2007.

[3] F. Darari, S. Razniewski, R. Prasojo, and W. Nutt. Enabling fine-grained RDF data completeness assessment. *ICWE*, 2016.

[4] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *SIGKDD*, 2014.

[5] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandecic. Introducing wikidata to the linked data web. In *ISWC*, 2014.

[6] L. Galárraga, C. Teflioudi, K. Hose, and F. Suchanek. Fast Rule Mining in Ontological Knowledge Bases with AMIE+. *VLDB Journal*, 2015.

[7] L. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek. AMIE: Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases . In *WWW*, 2013.

[8] A. Y. Levy. Obtaining complete answers from incomplete databases. In *VLDB*, 1996.

[9] B. Min, R. Grishman, L. Wan, C. Wang, and D. Gondek. Distant supervision for relation extraction with an incomplete knowledge base. In *HLT-NAACL*, pages 777–782, 2013.

[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 *AAAI*, 2015.

[11] A. Motro. Integrity = Validity + Completeness. *TODS*, 1989.

[12] F. Naumann. Data profiling revisited. *SIGMOD Record*, 2014.

[13] S. Razniewski, F. Korn, W. Nutt, and D. Srivastava. Identifying the extent of completeness of query answers over partially complete databases. In *SIGMOD*, 2015.

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

[15] B. Roth, T. Barth, M. Wiegand, and D. Klakow. A survey of noise reduction methods for distant supervision. In *AKBC*, pages 73–78. ACM, 2013.

[16] P. Spoor, M. Airey, C. Bennett, J. Greensill, and R. Williams. Use of the capture-recapture technique to evaluate the completeness of systematic literature searches. *BMJ*, 1996.

[17] F. M. Suchanek, S. Abiteboul, and P. Senellart. PARIS: Probabilistic Alignment of Relations, Instances, and Schema . In *VLDB*, 2012.

[18] F. M. Suchanek, D. Gross-Amblard, and S. Abiteboul. Watermarking for Ontologies . In *ISWC*, 2011.

[19] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *WWW*, 2007.

[20] S. Takamatsu, I. Sato, and H. Nakagawa. Reducing wrong labels in distant supervision for relation extraction. In *ACL*, pages 721–729. ACL, 2012.

[21] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 2014.