

Knowledge Base Recall: Detecting and Resolving the Unknown Unknowns

Simon Razniewski

and

Gerhard Weikum

Max Planck Institute for Informatics, Saarbrücken, Germany

Knowledge bases about entities like people, places and products have become key assets of search and recommender systems. The largest of them contain many millions of entities and billions of facts about them. Nevertheless, they have major gaps and limitations in what they cover, thus posing the challenge of detecting and resolving these “unknown unknowns”. This paper provides an overview on the problems in mapping knowledge base recall and the existing approaches to address these issues. Specifically, we discuss i) formalisms and tools for describing incompleteness, ii) rule mining methods to assess recall, iii) text mining methods to this end, and iv) approaches towards relative recall and informativeness.

1. PROBLEM DESCRIPTION

Knowledge bases, KBs for short, are structured collections of general world knowledge about entities (i.e., people, places, products, organizations and events), their properties and the relations between entities. KBs are a key asset for entity-centric search, question answering, entity linking and other tasks in NLP and AI applications. Companies such as Google, Microsoft, Alibaba or Bloomberg have in-house KBs, often referred to as knowledge graphs, as back-end infrastructure for search and recommender systems. Academic researchers and online communities offer publicly accessible KBs such as BabelNet [Navigli and Ponzetto 2012], DBpedia [Auer et al. 2007], Wikidata [Vrandečić and Krötzsch 2014] and YAGO [Suchanek et al. 2007], and have developed a suite of tools for KB construction, curation, querying and more. A widely used representation in all these works is *subject-predicate-object (SPO)* triples following the RDF data model.

The automatic construction of KBs has been greatly advanced in the last decade, with downstream applications such as entity-centric Internet search at Baidu, Bing, Google, etc., the IBM Watson system beating humans in a quiz show [IBM 2012], or systems that pass parts of 8th grade science tests [Schoenick et al. 2017]. However, despite these achievements, the *recall* or (in-)completeness of KBs is still poorly understood. While the correctness of SPO triples and the resulting precision of KBs is straightforward to evaluate via sampling, the extent to which KBs completely cover specific aspects is unclear and leads to “unknown unknowns”.

Example 1. Consider a KB that knows only one parent of Donald Trump and no children of his. The parent contained in the KB then is a known known, while the missing second parent is a known unknown. Whether the KB misses any children of Trump, in contrast, is an unknown unknown: by inspecting the KB alone, we cannot decide if he really has no children or if his children are merely missing in the KB.¹

Aside from a few well-behaved properties such as birthdate or birthplace, most predicates in KBs in fact lead to unknown unknowns, examples being spouses, children, books written, songs composed, awards won, etc. Assessing the recall of a KB is difficult for the following reasons:

- Representation: Data models for the Semantic Web operate under the open-world assumption, which does not consider completeness.
- Scale: KBs contain millions of subjects with thousands of predicates, making it infeasible to manually assess the recall of predicates.
- Evaluation: The usual quality studies based on fitness-for-use do not generalize beyond specific use cases.
- Skew: Although KBs tend to have high recall on popular subjects and predicates, they are overwhelmingly incomplete on the long tail of subjects and their SPO facts.

Understanding whether a KB has good recall on a topic is important both to consumers and creators (engineers/editors) of KBs: Knowing about recall helps consumers to better understand the reliability of data obtained, while for creators, it allows an informed decision on how to best allocate limited resources for curation.

Few researchers have looked at KB recall so far. To some extent, question answering evaluates KB recall as part of the evaluation of the whole pipeline. A paper by Mishra et al. has also looked at the specific role the KB played, finding that the best KB for science knowledge contains about 23% of facts needed for answering 4th grade science questions [Mishra et al. 2017]. But they also conceded that agreement was difficult to reach, and question answering is only one specific application for KBs.

In our research we aim to develop foundations and principled methods for assessing KB recall. In Section 2 we present the model and the formalisms we have developed for this purpose. In Sections 3 and 4 we discuss how data mining and text mining can be used for assessing *object recall*, i.e., whether all objects for a given SP-pair are present. Section 5 outlines an approach to *predicate recall*, i.e., assessing the coverage of predicates for a subject as a whole. Section 6 concludes the paper with open issues and further research opportunities.

2. FOUNDATIONS

Databases are classically interpreted under the Closed World Assumption (CWA): relational tuples in a database are assumed to be true facts, and tuples not in the database are assumed to be false statements [Minker 1982]. This is reasonable in limited domains such

¹Terminology made popular by Donald Rumsfeld in 2002.

as databases about employees, customers or warehouses. However, KBs aim to cover a much wider and potentially open-ended range of topics. In this setting, the natural paradigm is the Open World Assumption (OWA): information not contained in a KB is considered to be of unknown truth [McGuinness et al. 2004]. This is a safe approach insofar as it avoids incorrect conclusions; yet it does not do full justice to KBs. On many topics, KBs are practically complete, implying that information not contained in a KB is indeed not true. This observation motivates an intermediate ground: the *partially closed world assumption* (PCWA). The PCWA generally adopts the OWA, but uses *completeness statements* to specify parts of its data that should be treated under closed world semantics [Motro 1989; Levy 1996].

Formally, completeness statements are constraints relating the real world and the data that describes it. While the real world is generally open-ended, completeness expresses that in a specific part, the data at hand fully covers what is true in reality. A model-theoretic formalization is given in [Razniewski et al. 2015], where completeness statements are formalized as restricting the set of states the real world can be in. For languages of completeness statements, ranging from conjunctive-query fragments to simple selections, an important issue is to determine the completeness (or potential incompleteness) of query results from the partial completeness of the data. This inference problem has been investigated in [Dencker et al. 2010; Razniewski and Nutt 2011; Darari et al. 2013].

For KBs, an appropriate language of completeness statements are *SP-statements* [Darari et al. 2016]. SP-statements are pairs of a subject S and a predicate P and express that for subject S, all objects for predicate P are known.

Example 2. With an SP-statement for the predicate *brother* of subject *Trump*, we can express that Wikidata contains all brothers of Trump. Consequently, anybody not explicitly stated to be a brother of Trump, such as Barack Obama, is not Trump’s brother.

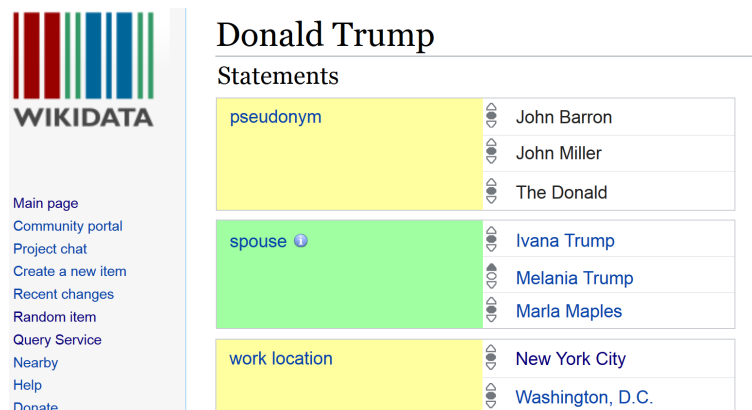
For communities that maintain KBs, such as Wikidata, it is in the interests of editors to understand the complete and incomplete parts of the KB. We developed the COOL-WD² [Darari et al. 2017] tool (*Completeness tool for Wikidata*), that allows editors to record complete SP-pairs in Wikidata. COOL-WD adds an overlay to the default Wikidata interface that shows SP-statements, such as the green (dark) box around the predicate *spouse* for *Trump* in Fig. 1. The SP-statements can then be used in query answering, for instance to check whether a query for all spouses of US presidents gives a complete answer. Another use case is profiling the completeness of sets of SP-pairs, for instance, sorting states by completeness with regard to their counties.

However, relying on manual assertions by editors is inherently limited in its scale. Therefore, the following sections describe directions we are pursuing towards the automatic assessment of KB recall.

3. ASSESSING KB RECALL USING RULE MINING

Algorithms for association rule mining have been applied to KBs to derive deduction rules and predict missing facts [Galárraga et al. 2015]. It is conceivable that such approaches

²<http://cool-wd.inf.unibz.it>



The screenshot shows the Wikidata page for Donald Trump. On the left is the Wikidata navigation menu. The main content area is titled 'Donald Trump' and 'Statements'. It lists three categories of statements, each with a colored background and a completeness icon:

- pseudonym** (yellow background, icon with 2 dots): John Barron, John Miller, The Donald.
- spouse** (green background, icon with 3 dots): Ivana Trump, Melania Trump, Marla Maples.
- work location** (yellow background, icon with 2 dots): New York City, Washington, D.C.

Fig. 1. Recall information inside Wikidata as enabled by the COOL-WD tool. The green (darker) *spouse* predicate is asserted to be complete, while for *pseudonym* and *work location* completeness is unknown.

can also reveal patterns about the (in-)completeness of KBs. For instance, people having a death place but no death date are probably missing the latter, while people having two parents are likely complete regarding parents. A more intricate pattern that can be found this way is that movies with a known producer are usually complete on directors (the reason being that the latter information is usually more important than the former).

All these examples can be expressed as logical rules:

Example 3.

- $hasPlaceOfDeath(x) \wedge hasNoDateOfDeath(x) \Rightarrow incomplete(x, dateOfDeath)$
- $hasTwoParents(x) \Rightarrow complete(x, parents)$
- $hasProducer(x) \Rightarrow complete(x, director)$

In [Galárraga et al. 2017] we investigated how such rules can be automatically mined. We designed a rule language that includes two new target predicates about (in-)completeness – *complete* and *incomplete* (e.g., $complete(x, parents)$), predicates about object counts (e.g., $hasTwoParents(x)$), and a limited form of negation in a way that ensured that all rules could be expressed as Horn rules [Horn 1951]. With the manual insertion of instances of these target predicates, we could then approach the pattern mining for (in-)completeness as a Horn rule learning problem. To this end, we extended the AMIE system [Galárraga et al. 2015] with the new predicates, and used crowdsourcing to label 3,000 subject-predicate-pairs with *complete* or *incomplete* predicates. The extended AMIE system could then predict new instances of target predicates for a variety of Wikidata properties with an F1 score of 69% to 100%. Table I shows results for a subset of these properties. The results show that patterns in data can give valuable cues about the recall of KBs.

As another extension, we have shown how rules about cardinality constraints can be learned using rule mining [Tanon et al. 2017]. An example is that people have at least as many children as their children have siblings. Such constraints can then be used to infer how many facts a KB should contain for certain SP-pairs.

Two important questions left open for future work are (i) the handling of conflicting pre-

Relation	F1 Score
alma mater	87%
brother	96%
child	73%
country of citizenship	98%
director	100%
father	100%
mother	100%
place of birth	100%
place of death	96%
sex or gender	100%
spouse	55%

Table I. F1 score of the extended AMIE system on predicting completeness/incompleteness for selected Wikidata properties.

dictions, and (ii) how to create instances of the target predicates. Regarding conflicting predictions, note that association rules were originally developed for settings where patterns in complete data are to be found. In the case of incomplete data, association rules can lead to conflicting predictions, for instance that a person is born in France and in Sweden, or that a predicate of a person is complete and incomplete. Present approaches such as AMIE assume statistical independence of rules to resolve these conflicts.

The second issue concerns the automatic generation of target predicate instances. To mine patterns for cues about (in-)completeness, a sufficient number of instances is required. Current KBs do not contain these (meta-)predicates at all, and instances cannot be generated from the KBs themselves. For example, even when a KB contains three (ex-)spouses of Trump, this does not tell us whether these are all spouses of his. We investigated two heuristics for automatically generating target predicates: (i) assuming that subjects that stand in a relation with at least one object are complete for that relation (so-called partial completeness assumption [Galárraga et al. 2015], an instance of the PCWA), and (ii) assuming that subjects of sufficient popularity are complete. However, both heuristics have limitations. Thus, better methods for automated target predicate generation are called for.

4. ASSESSING KB RECALL USING TEXT EXTRACTION

Information extraction (IE) from text is a common methodology for KB construction, and texts are useful resources even for KBs that are constructed by other means. We are therefore investigating to which extent IE can help in KB recall assessment. Texts often give strong cues on how many objects stand in a certain relation with a certain subject, as the following examples show.

Example 4.

- (1) The city consists of seven boroughs.
- (2) She was married twice.
- (3) This TV series has 5 episodes.
- (4) They have no children.
- (5) The team has few midfielders.

The first sentence, for instance, tells that the knowledge base should contain seven boroughs that stand in the relation *partOf*. We collectively refer to such cues as *cardinality cues*, as they give information on the cardinality of the set of objects standing in relation with a subject. So far we have focused on cardinality cues that use explicit numerals (example sentences 1, 2 and 3). By extracting that the number *seven* above expresses the number of boroughs for a given city, we can look up the number of such facts in the KB, and determine the completeness or incompleteness for this city.

Interestingly, these numeric cues have been disregarded by most IE research. We found that 19% of numbers in Wikipedia articles express the cardinality of objects that stand in a relation with a certain subject. So far, IE methods only deal with numbers in the context of time and quantities such as money or physical measures (see e.g., [Strötgen and Gertz 2010; Chaganty and Liang 2016]).

In [Mirza et al. 2016], we showed that a small number of manually defined patterns can extract counting information that implies the existence of 178% more *hasChild* triples than Wikidata currently knows of. By comparing expected children counts with actual children counts for people of different types or professions, we can get insight into the completeness on certain topical slices of the KB. For instance, we found children count information for 8.22% of all judges and 5.11% of all politicians, and matches between textually extracted counts and Wikidata triples for only 2.79% of all film directors and 0.13% of all baseball players. In [Mirza et al. 2017], we developed a cardinality IE framework based on a Conditional Random Field for sequence labeling, and achieved 32%-55% precision in extracting counts for the *child*, *administrative subdivision* and *part of creative works* predicates.

This research direction is still in its infancy. As shown in Example 4, there are several ways to express cardinalities other than via explicit numerals. In particular, some cues require more complex inference, for instance when they only express bounds (“*Her second spouse*”) or are compositional, like in “*two children from the first marriage and one from the second*”. Addressing such cases requires new ways of reasoning.

5. PREDICATE RECALL

The methods in the previous two sections aim to determine recall regarding specific properties of subjects, i.e., SP-pairs, like (ex-)spouses of Trump or districts of a city. Yet information needs are not always that specific; it is also important to assess the completeness of information about a subject S as a whole. How many distinct predicates that should be present are indeed covered by the KB?

A logical notion of completeness is difficult to devise in this case for the lack of a well-defined reference point. Even with a fixed set of pre-defined predicates the problem is challenging, due to the large number of candidate predicates (e.g., ca. 3,000 in Wikidata) and the sometimes fuzzy meaning of certain predicates (e.g., *influencedBy*). To quantify completeness under these circumstances, we propose the use of a *relative* notion of recall: capturing recall in comparison with other, similar subjects. For example, to assess the completeness of data about Trump, one should look at the KB contents for other US presidents, and for assessing the completeness of a city, one should look at the data for similar cities.

Formally, relative recall relies on two components:

The image shows a Wikidata page for Donald Knuth. On the left is the Wikidata logo and navigation links: Main page, Community portal, Project chat, Create a new item, and Recent changes. The main content area displays the name 'Donald Knuth' with a small green status indicator in the top right corner. Below the name is a section titled 'Most relevant properties which are absent' with a dropdown arrow. It lists five properties with their respective percentages: P734 - family name - 12.48%, P937 - work location - 12.04%, P103 - native language - 4.72%, P102 - member of political party - 4.7%, and P136 - genre - 3.64%. Below this list is a table showing the 'occupation' property with two values: 'mathematician' and 'computer scientist', each with a small green status indicator.

Fig. 2. Relative recall information in Wikidata as enabled by Recoin. The plugin adds a status indicator (top-right corner) and a list of important missing properties (center). Numbers behind missing properties indicate the percentage of similar subjects that have objects for these properties.

- (1) A similarity function between subject pairs $sim(S_1, S_2)$ that can be used to compute a (weighted) set of similar subjects \mathbf{S} for a subject S .
- (2) A scoring function $score(S, \mathbf{S})$ that computes a score or rank for the completeness of S with regard to a set of comparison subjects \mathbf{S} .

We have devised such a relative recall model for Wikidata, and deployed a tool called Recoin [Ahmeti et al. 2017] (Relative completeness indicator)³. Recoin uses a simple Boolean similarity function that considers two people as similar if they share at least one profession, and subjects of other types (i.e., locations, organizations, etc.) as similar if they share at least one (non-trivial) type. The tool then computes the 5 most common properties in \mathbf{S} that subject S is missing, and shows the average frequency of these properties in the comparison set \mathbf{S} as $score(S, \mathbf{S})$.

An example is shown in Fig. 2. Donald Knuth is compared with all humans that have profession *mathematician* or *computer scientist*. The most frequent properties among these subjects that Knuth is lacking are *work location*, *family name*, *native language*, *political party* and *genre*, which occur in 12.48% to 3.64% of these types of subjects, thus leading to an incompleteness score of 7.52. For comparison, Trump’s score is 1.93, while Tim Berners-Lee’s score is 4.37.

Both similarity and scoring function leave space for refinement. For computing subject similarity, a range of techniques such as similarity of textual descriptions or relatedness measures in knowledge graphs could be used (for a recent survey, see [Ponza et al. 2017]). For scoring subjects, it is desirable to use more informed techniques than simple counts, as frequent properties are not necessarily also important. In our ongoing work, we aim to devise more accurate and subject-specific rankings of properties [Razniewski et al. 2017].

³<https://www.wikidata.org/wiki/Wikidata:Recoin>

6. OPEN CHALLENGES

The presented approaches are steps towards a principled understanding of KB recall. Open challenges along these directions include the following.

- Predicate salience*: While subject salience [Hachey et al. 2013] and fact salience [Bast et al. 2015] have received a fair amount of attention, assessing if a subjects’ set of facts is complete requires a notion of predicate salience. For example, the predicate *wroteBook* is crucial for a writer, and would also be informative for say a football player. In contrast, the predicate *drivesCarModel* is of little interest for most people. To date, algorithms are performing considerably worse than humans in judging predicate salience (e.g., 75% versus 87.5% precision in pairwise preference as reported in [Razniewski et al. 2017]).
- Relative completeness*: Our approach to quantifying predicate completeness, Recoin, can be extended both by more refined subject similarity metrics and scoring methods. Also, human-generated assessments of relative completeness do not exist so far, and the design of benchmarks would be valuable.
- Training data*: Rule mining and extraction from text require KB facts as seeds for pattern mining or distantly supervised learning. Learning to assess incompleteness is different from acquiring new facts, though. In our setting, we also need incomplete seeds of different kinds – a special and subtle case of negative training data.

REFERENCES

- AHMETI, A., RAZNIEWSKI, S., AND POLLERES, A. 2017. Assessing the completeness of entities in knowledge bases. *14th Extended Semantic Web Conference (ESWC) Posters & Demos*.
- AUER, S., BIZER, C., KOBILAROV, G., LEHMANN, J., CYGANIAK, R., AND IVES, Z. G. 2007. DBpedia: A nucleus for a web of open data. *6th International Semantic Web Conference (ISWC)*, 722–735.
- BAST, H., BUCHHOLD, B., AND HAUSSMANN, E. 2015. Relevance scores for triples from type-like relations. *38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 243–252.
- CHAGANTY, A. T. AND LIANG, P. 2016. How much is 131 million dollars? putting numbers in perspective with compositional descriptions. *54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 578–587.
- DARARI, F., NUTT, W., PIRRÒ, G., AND RAZNIEWSKI, S. 2013. Completeness statements about RDF data sources and their use for query answering. *12th International Semantic Web Conference (ISWC)*, 66–83.
- DARARI, F., PRASOJO, R. E., RAZNIEWSKI, S., AND NUTT, W. 2017. COOL-WD: A completeness tool for Wikidata. *16th International Semantic Web Conference (ISWC) Posters & Demos*.
- DARARI, F., RAZNIEWSKI, S., PRASOJO, R. E., AND NUTT, W. 2016. Enabling fine-grained RDF data completeness assessment. *International Conference on Web Engineering*, 170–187.
- DENECKER, M., CORTÉS-CALABUIG, A., BRUYNNOGHES, M., AND ARIELI, O. 2010. Towards a logical reconstruction of a theory for locally closed databases. *ACM Transactions on Database Systems* 35, 3.
- GALÁRRAGA, L., RAZNIEWSKI, S., AMARILLI, A., AND SUCHANEK, F. M. 2017. Predicting completeness in knowledge bases. *10th ACM International Conference on Web Search and Data Mining (WSDM)*.
- GALÁRRAGA, L., TEFLIOUDI, C., HOSE, K., AND SUCHANEK, F. M. 2015. Fast rule mining in ontological knowledge bases with AMIE+. *VLDB Journal* 24, 6, 707–730.
- HACHEY, B., RADFORD, W., NOTHMAN, J., HONNIBAL, M., AND CURRAN, J. R. 2013. Evaluating entity linking with Wikipedia. *Artificial Intelligence* 194, 130 – 150.
- HORN, A. 1951. On sentences which are true of direct unions of algebras. *The Journal of Symbolic Logic* 16, 1, 14–21.
- IBM. 2012. Special issue on “This is Watson”. *IBM Journal of Research and Development* 56, 3.
- SIGWEB Newsletter Winter 2017

- LEVY, A. Y. 1996. Obtaining complete answers from incomplete databases. *22nd International Conference on Very Large Data Bases (VLDB)*, 402–412.
- MCGUINNESS, D. L., VAN HARMELEN, F., ET AL. 2004. OWL web ontology language overview. *W3C recommendation 10*, 10.
- MINKER, J. 1982. On indefinite databases and the closed world assumption. *6th Conference on Automated Deduction*, 292–308.
- MIRZA, P., RAZNIEWSKI, S., DARARI, F., AND WEIKUM, G. 2017. Cardinal virtues: Extracting relation cardinalities from text. *55th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- MIRZA, P., RAZNIEWSKI, S., AND NUTT, W. 2016. Expanding Wikidata’s parenthood information by 178%, or how to mine relation cardinality information. *15th International Semantic Web Conference (ISWC) Posters & Demos*.
- MISHRA, B. D., TANDON, N., AND CLARK, P. 2017. Domain-targeted, high precision knowledge extraction. *Transactions of the Association for Computational Linguistics 5*, 233–246.
- MOTRO, A. 1989. Integrity = Validity + Completeness. *ACM Transactions on Database Systems 14*, 4, 480–502.
- NAVIGLI, R. AND PONZETTO, S. P. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence 193*, 217–250.
- PONZA, M., FERRAGINA, P., AND CHAKRABARTI, S. 2017. A two-stage framework for computing entity relatedness in Wikipedia. *26th ACM International Conference on Information and Knowledge Management (CIKM)*, 1867–1876.
- RAZNIEWSKI, S., BALARAMAN, V., AND NUTT, W. 2017. Doctoral advisor or medical condition: Towards entity-specific rankings of knowledge base properties. *13th International Conference on Advanced Data Mining and Applications (ADMA)*.
- RAZNIEWSKI, S., KORN, F., NUTT, W., AND SRIVASTAVA, D. 2015. Identifying the extent of completeness of query answers over partially complete databases. *ACM SIGMOD International Conference on Management of Data*, 561–576.
- RAZNIEWSKI, S. AND NUTT, W. 2011. Completeness of queries over incomplete databases. *37th International Conference on Very Large Data Bases (VLDB) 4*, 11, 749–760.
- SCHOENICK, C., CLARK, P., TAFJORD, O., TURNEY, P., AND ETZIONI, O. 2017. Moving beyond the Turing Test with the Allen AI Science Challenge. *Communications of the ACM 60*, 9 (8), 60–64.
- STRÖTGEN, J. AND GERTZ, M. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. *5th International Workshop on Semantic Evaluation*, 321–324.
- SUCHANEK, F. M., KASNECI, G., AND WEIKUM, G. 2007. YAGO: a core of semantic knowledge. *16th International Conference on World Wide Web (WWW)*, 697–706.
- TANON, T. P., STEPANOVA, D., RAZNIEWSKI, S., MIRZA, P., AND WEIKUM, G. 2017. Completeness-aware rule learning from knowledge graphs. *16th International Semantic Web Conference (ISWC)*.
- VRANDEČIĆ, D. AND KRÖTZSCH, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM 57*, 10, 78–85.

Simon Razniewski is researcher at the Max Planck Institute for Informatics, and was previously Assistant Professor at the Free University of Bozen-Bolzano. His research spans the foundations of database management and data quality, and has been published at conferences such as VLDB, SIGMOD, ACL, CIKM and ISWC.

Gerhard Weikum is a Scientific Director at the Max Planck Institute for Informatics, where he is leading the department on databases and information systems. His research spans transactional and distributed systems, self-tuning database systems, data and text integration, and the automatic construction of knowledge bases. He co-authored a comprehensive textbook on transactional systems, received the VLDB 10-Year Award for his work on automatic DB tuning, and is one of the creators of the YAGO knowledge base.