# Structured knowledge: Have we made progress? An extrinsic study of KB coverage over 19 years

Simon Razniewski and Priyanka Das
Max Planck Institute for Informatics
srazniew@mpi-inf.mpg.de

## ABSTRACT

Structured world knowledge is at the foundation of knowledge-centric AI applications. Despite considerable research on knowledge base construction, beyond mere statement counts, little is known about the progress of KBs, in particular concerning their coverage, and one may wonder whether there is constant progress, or diminishing returns. In this paper we employ question answering and entity summarization as extrinsic use cases for a longitudinal study of the progress of KB coverage. Our analysis shows a near-continuous improvement of two popular KBs, DBpedia and Wikidata, over the last 19 years, with little signs of flattening out or leveling off.

**ACM Reference Format:**
Simon Razniewski and Priyanka Das. 2020. Structured knowledge: Have we made progress? An extrinsic study of KB coverage over 19 years. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3417447

## 1 PROBLEM OVERVIEW

**Motivation and problem.** Knowledge bases (KBs), that is, structured and machine-readable collections of world knowledge, have experienced significant growth in recent years. From their inception in 2006 and 2012, respectively, DBpedia [3] and Wikidata [33] have grown from an initial 102 million triples to over 3 billion billion[1], and 26 million statements at the end of 2013 to nearly 1 billion statements as of today[2]. A similar growth can also be conjectured for industrial projects, as pursued, e.g., at Google, Microsoft or Amazon [23]. Yet these numbers do not tell how growth in size relates to growth in *coverage*. Is the size growth mirrored by a corresponding increase in the coverage of *relevant* information? Or does the growth merely result from adding irrelevant iotas and dots? What are these KBs still missing?

Knowledge base construction has received considerable attention in the literature, with works often competing for bigger and bigger KBs (see e.g., the billion triple challenge [19] or Table 1 in [11]), and precision often evaluated by sampling [30]. About their recall, by comparison, much less is known. Several projects have evaluated recall of specific topics, using either KB-intrinsic statistical methods [13, 18], or external evidence from texts [20, 27]. Yet these approaches lack a notion of *relevance*. They can estimate whether a KB likely knows all children of President Trump, or all books by Stephen King, but do not quantify the relevance of these. Closest to our problem is work in [14], which uses proprietary Amazon Alexa query logs in order to identify the relevance of properties for entities, though this is subsequently used to suggest priorities in knowledge acquisition, not to study KB coverage and its temporal dynamics.

Mapping trajectories is crucial to understand the dynamics of knowledge base construction field. Are we on a steady upward trajectory? Was the major impact achieved early on, and since then we are only polishing bits and pieces? Were there major steps in the development? There are several proposals to practically extend existing knowledge representation formalisms for web-scale knowledge bases [2, 29, 31], are they hinting at relevant problems?

**Approach and contribution.** To enable discussions and direct further KB research, in this paper we present a longitudinal study of KB coverage. Our study relies on two crucial components:

(1) We use question answering and entity summarization to ground relevance of information.
(2) We exploit historical KB data in order to analyze their change in coverage.

To our knowledge, this is the first longitudinal study of KB coverage, and it closes a critical gap of previous faithful research on knowledge bases. Remarkably, our study finds no indication of diminishing returns or a levelling-off of KB usefulness, thus indicating that previous size improvements have indeed been mirrored in value improvements, and one can conjecture that in the near future, this trend will continue.

## 2 METHODOLOGY

**Knowledge bases.** We utilize two popular open knowledge bases, DBpedia [3] and Wikidata [33].

(1) *DBpedia* is built by scraping infoboxes from Wikipedia, and started in 2006, presents one of the earliest efforts at web-scale KB construction. By inspecting earlier versions of Wikipedia pages, we can simulate the content of the DBpedia knowledge base since the start of Wikipedia in 2001.
(2) *Wikidata* is a community-built KB with an edit history ranging back until 2012. Following the now-defunct Freebase, it represents currently the most comprehensive effort at collaborative KB construction.

While Wikidata has a detailed edit history available, DBpedia dumps were only published sporadically (e.g., no dumps after 2016). To

[1] https://wiki.dbpedia.org/about/facts-figures
[2] https://tools.wmflabs.org/wikidata-todo/stats.php

*CIKM '20, October 19–23, 2020, Virtual Event, Ireland*

ACM ISBN 978-1-4503-6859-9/20/10.
https://doi.org/10.1145/3340531.3417447

| Query | First answerable |
|---|---|
| *how old is dustin pedroia* | May 18, 2017 |
| *where is italian job filmed* | October 15, 2015 |
| *what type of government is ontario* | April 22, 2020 |
| *what time zone is ohio in* | August 31, 2013 |

**Table 1: Sample queries from Google Suggest, along with earliest date they could be answered by Wikidata.**

obtain dense information, we thus simulate the potential DBpedia content at specific dates by inspecting Wikipedia's infobox content, for which continuous data is available via its edit history.

**Question preparation.** We utilize the following three sources of questions:

(1) *Google Suggest*: 1 million query suggestions from the Google web search engine, collected via repeated (breadth-first) gathering of related queries [5].[3]
(2) *AOL queries*: 20 million web search queries from AOL, collected in 2006 [22].[4]
(3) *MS Marco*: 1 million Bing queries from the MS MARCO dataset [21].

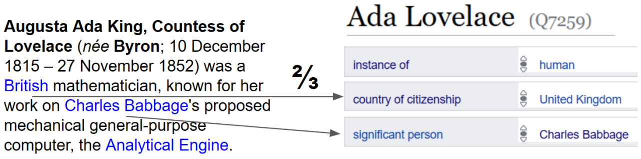
The three sources come with different characteristics. The Google Suggest queries are already almost entirely questions (i.e., phrases starting with a question word), whereas the AOL query log is composed mostly of keyword queries that are not easily interpretable (e.g., *"Apple keyboard"*, *"Houston sights"*) and not in our focus. Beyond that, there are various other kinds of issues for the present analysis, such as personal pronouns, and relative and absolute temporal markers. We thus used automated filters to remove the following classes of queries:

(1) Queries not starting with a question word (*"lyrics who let the dogs out"*).
(2) Queries w/o named entities (*"where can i buy alcohol")*
(3) Queries that contain relative temporal markers (*'Who is pitching for the Tigers tomorrow"*).
(4) Queries containing *how to/how can* (*"how can one join NASA"*)
(5) Queries containing personal pronouns (*"what is recommended insurance for me"*)

From samples used for annotation, we also manually removed those that had no answer in 2001 (e.g., *"Who did Trump compete against"*), as for these, an inability to answer them initially is obviously not a shortcoming of KBs. Examples of surviving queries, and the first time they could be answered on Wikidata, are shown in Table 1.

**Question annotation scheme.** From each filtered query source, we randomly sampled 100 queries, and used expert annotation to determine whether it could be answered by each KB, and since when. As working with the revision history of both Wikipedia and Wikidata is unwieldy, archived snapshots from the internet archive (archive.org/web) were used as the basis of the annotations.

Deciding on KB-answerability required advanced training on knowledge representation and query answering, we thus relied on one expert annotator to perform the annotations. The annotator spent a total of about 48 hours (5 minutes times 100 queries times 3 query datasets times 2 KBs) on the task.


**Augusta Ada King, Countess of Lovelace** (*née* **Byron**; 10 December 1815 – 27 November 1852) was a British mathematician, known for her work on Charles Babbage's proposed mechanical general-purpose computer, the Analytical Engine.

Ada Lovelace (Q7259)

| | |
|---|---|
| instance of | human |
| country of citizenship | United Kingdom |
| significant person | Charles Babbage |

**Figure 1: Example of entity summarization.**

The annotation task was split into three steps:

(1) Given a query, decide whether its intent is sufficiently unambiguous (positive example: *When was Lincoln born?* Negative example: *What is the best place in Texas?*)
(2) If its semantics are sufficiently unambiguous, determine whether the query can be answered by a KB as of May 2020, by spending at most 5 minutes searching for and browsing through relevant fact pages.
(3) If the query could be answered in 2020, then determine the earliest date it could be answered by performing binary search on the history of the page(s) that contained the answer.

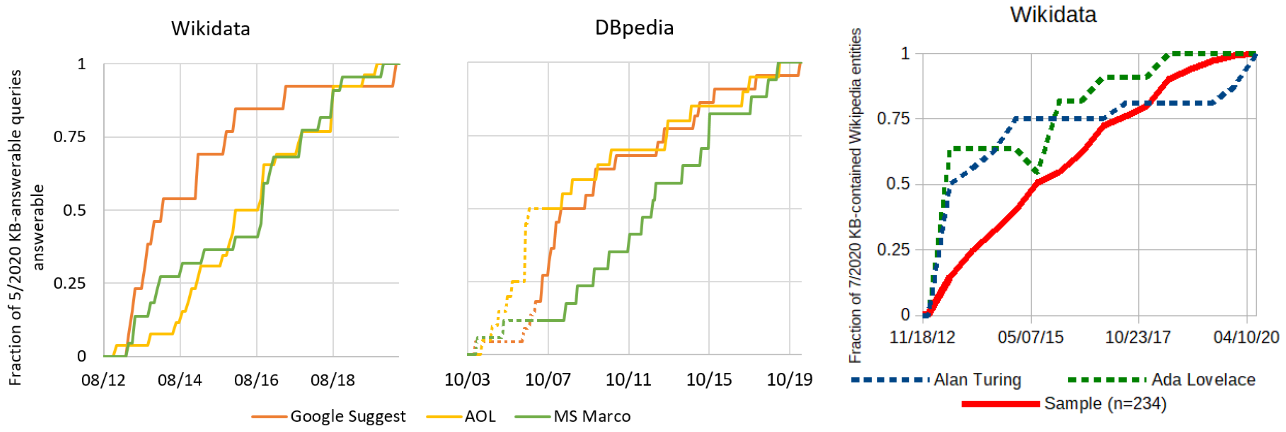
The annotator was instructed to be tolerant to reformulations and paraphrases (e.g., Query: *What are popular foods in Spain?* - KB content: *(Paella, national food of, Spain)*, and to accept partial answers (e.g., there is no notion of a complete answer to the above query, but if the KB contained at least one of two popular responses, then this was deemed sufficient). The full annotation guidelines are available at https://tinyurl.com/kb-coverage-guidelines.

**Entity summarization.** Our second extrinsic yardstick is entity summarization - how well a KB captures salient knowledge about an entity. We employed Wikipedia articles, where we focus on facts concerning the most important entities, syntactically distinguished by link markup. We employ the distant supervision assumption that a fact is represented where its entities are present, i.e., do not require matching predicates. In other words, we check how many of the entities linked in the Wikipedia article of a subject are also present as objects in KB triples for the subject, and how that number changes over time (see Fig. 1). To avoid effects resulting from an evolving reality, we utilize snapshots of Wikipedia articles from 2012 as reference. Also, as DBpedia's content is too close to Wikipedia, we only perform this analysis for Wikidata, selecting a reproducible sample of 234 entities (all persons whose name contains "John Smith") with an English Wikipedia article.

## 3 RESULTS AND DISCUSSION

**Results.** The main results are shown in Figure 2. In (a) we show for each KB and query log the fraction of queries that could be answered at each timepoint, relative to all that could be answered in 5/2020, based on an absolute number of 13-26 answerable queries, out of a sample of 100 per dataset. Data points for the DBpedia extraction framework before its actual inception in 2006 are dashed. Similarly, in (b) we show the fraction of Wikipedia-linked entities present in Wikidata, relative to the total as of 2020, for a sample of 234 subjects, and 2 examples.

[3] The popular WebQuestions dataset [5] contains 6642 of these, which represent the fraction of a sample of 100k that in 2012 could be answered by Freebase, but the authors kindly provided the full dataset.
[4] This dataset has been surrounded by controversies, as it contains session data that has allowed identification of individual users. We discarded all session data.



**(a) Number of KB-answerable queries over time.**

**(b) Number of KB-contained summary entities over time.**

**Figure 2: Main results.**

As one can see, the growth is largely evenly spread, with only minor growth spurts of individual curves in shorter intervals, e.g., AOL queries on the DBpedia data in 2006. The faster growth of Google Suggest may be attributed to its origin as sanitized recommendation, i.e., a set of likely more popular queries. Importantly, we do not observe a systematic ceiling effect or levelling off of the curves.

**Relative vs. absolute numbers.** Furthermore, it is worth noting that worldwide information access, search engine queries, and question answering (especially on mobile devices) have been significantly increasing in the last years, thus, in *absolute* terms, moderate decreases in the relative growth of answerable queries are easily offset, and the growth in the absolute number queries that are KB answerable is quite likely still increasing.[5]

**Representativeness of results.** DBpedia and Wikidata are exemplary for KBs built by collaborative manual effort, a paradigm also pursued by Yago and Freebase. Moreover, Wikidata more recently embraced large-scale data integration, importing for example 300k members of the British peerage[6], and represents to our knowledge the best public equivalent of closed industrial efforts with heavy focus on source selection and integration [34].

**Progress drivers.** An important question is to understand what the drivers of this progress were. Although a detailed analysis is beyond the scope of this paper, we can relate the number of answerable queries with other metrics, in particular, the number of statements, and active editors. Wikidata editors, for example, have seen a near-linear growth from 4,000 active editors in 2013, to 12,000 in 2020.[7] Similarly the number of Wikidata entities and statements has grown in a linear fashion, with marked changes of the derivative, to up to 86 million entities and 1 billion statements in 2020.

| | DBpedia | Wikidata |
|---|---|---|
| Main entity in KB | 58% | 87% |
| Main entity has data in KB | 58% | 75% |
| Query KB-answerable for another entity | 35% | 35% |
| Query answerable from Wikipedia text | 47% | 48% |

**Table 2: Statistics on unanswerable questions (n=120 - 20 per query log per KB).**

The English Wikipedia, in contrast, has remarkedly observed a decline of active editors, from over 60,000 in 2007, to 44,000 in 2020. Nonetheless, its number of articles observed a roughly linear growth (from 500k in 2005 to over 6 million in 2020), and its content size has grown similarly linearly, indicating that the growth in the number of articles not just represents an addition of stubs, but an overall organic growth of content.[8]

Qualitatively, we can observe that infoboxes (DBpedia) have undergone significant standardization over the years, while Wikidata has observed a significant integration of external data. More insights into qualitative drivers of KB growth are contained in [35].

**The future.** *"It is difficult to make predictions, especially about the future."* Nonetheless, the observed trends are a reason for optimism, and we conjecture that KBs coverage will continue to grow in the near future. We can not predict how progress will be achieved, but a naive indication is given by statistics around questions that currently could *not* be answered by KBs, which we show in Table 2.

## 4 RELATED WORK

**Knowledge base construction.** Knowledge base construction has a long history in AI and data management [17]. The field picked up pace in the 2000s with the DBpedia [3] and Yago [30] projects that harvested knowledge from Wikipedia, and web extraction efforts like Nell [7] and KnowItAll [12]. The usefulness is proven by the widespread deployment of in-house KBs at most major tech companies [23], powering and supporting applications like entity

[5] As search engine traffic is a well-guarded secret an analysis is beyond our means, although the few available cues hint at a moderate polynomial or even exponential growth (see e.g., https://www.internetlivestats.com/google-search-statistics).

[6] https://www.wikitree.com/g2g/951698/bot-import-of-the-peerage-data-in-wikidata-good-or-bad-news

[7] https://stats.wikimedia.org/#/wikidata.org/contributing/active-editors/normal|line|all|~total|monthly

[8] https://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia%27s_growth

resolution, search and question answering, and by a vibrant startup scene around KBC. Nonetheless, editor biases may lead to systematic coverage gaps [9], and so far, little is known about the dynamics of the usefulness of KBs.

**KB recall assessment.** While precision is long in the focus of KBC, recall has only become a focus more recently [15, 16]. Now several proposals have been made in order to estimate the recall of KBs, in particular association rule mining [13], entity-entity-comparison [4], sampling-based extrapolation [18], and verification via external textual evidence [20, 27]. Yet intrinsic recall is not necessarily linearly related to extrinsic coverage, e.g., there might be a major growth in extrinsic coverage when increasing recall from 10% to 20%, but none when moving from 70% to 80%.

**Query logs and other relevance data.** Query logs are a prime source for understanding user information needs. Multiple web search query logs are available online, often with the intention to allow studying query features [6], or to evaluate the performance of question answering systems [5]. An interesting exception are experiments at Amazon that used non-disclosed Alexa query logs in order to understand which fields of entities should be completed first [14]. Related studies have also been performed in the field of commonsense knowledge acquisition, where the TupleKB project used textbook statements to estimate recall of relevant knowledge [8], and the Quasimodo project used crowd-generated knowledge [28]. Nonetheless, no study has been conducted on the temporal trajectory of KB coverage.

**Question answering and summarization.** QA and KBQA comprise another major research field, with much interest in improving methods that understand questions, traverse knowledge repositories, and identify and rank answer candidates. Progress in the field can be tracked by popular benchmarks like Free917 and WebQuestions, which consist of question-answer pairs with verified answers over a knowledge base [37]. For example, on WebQuestions, the F1-score went up from 36% to 56% in the years 2013-2016.[9] For summarization, approaches typically focus on content selection and ranking [1], or text generation. In the present paper, we take a very different approach, measuring not the progress of QA and summarization *systems*, but of underlying *structured data*.

**Other related topics.** Related to any question of coverage are cost-benefit tradeoffs. To this end, the cost of KB construction has been estimated in [24], and the cost of keeping them up-to-date in [26]. There are various proposals to increase the expressivity of KBs, in particular towards beliefs and opinions [29], negative knowledge [2], and arbitrary functions [31, 32]. A comprehensive qualitative analysis of lessons learned in KB construction is also contained in [35]. Historical KB data has been used in a few other contexts, in particular in the prediction of change [10, 36], and in an analysis of the stability of KB schemas [25].

**Acknowledgments.** We are thankful to Jianfeng Gao for discussions that triggered this study, to Paramita for help in coding, and to Jonathan Berant for providing the full Google Suggest dataset.

## REFERENCES

[1] Tim Althoff, Xin Luna Dong, Kevin Murphy, Safa Alai, Van Dang, and Wei Zhang. TimeMachine: Timeline generation for knowledge-base entities. In *KDD*, 2015.

[2] Hiba Arnaout, Simon Razniewski, and Gerhard Weikum. Enriching Knowledge Bases with Interesting Negative Statements. *AKBC*, 2020.

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a Web of open data. *ISWC*, 2007.

[4] Vevake Balaraman, Simon Razniewski, and Werner Nutt. Recoin: Relative Completeness in Wikidata. *Wiki workshop at WWW*, 2018.

[5] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, 2013.

[6] David J. Brenes and Daniel Gayo-Avello. Stratified analysis of AOL query log. *Information Sciences*, 2009.

[7] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. Toward an Architecture for Never-Ending Language Learning. *AAAI*, 2010.

[8] Bhavana Dalvi, Niket Tandon, and Peter Clark. Domain-Targeted, High Precision Knowledge Extraction. *TACL*, 2017.

[9] Gianluca Demartini. Implicit Bias in Crowdsourced Knowledge Graphs. In *Companion of WWW*, 2019.

[10] Ioannis Dikeoulias, Jannik Strötgen, and Simon Razniewski. Epitaph or Breaking News? Analyzing and Predicting the Stability of Knowledge Base Properties. *TempWeb workshop at WWW*, 2019.

[11] Xin Dong and others. Knowledge Vault: A web-scale approach to probabilistic knowledge fusion. *KDD*, 2014.

[12] Oren Etzioni and others. Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 2005.

[13] Luis Galárraga, Simon Razniewski, Antoine Amarilli, and Fabian M Suchanek. Predicting Completeness in Knowledge Bases. In *WSDM*, 2017.

[14] Andrew Hopkinson, Amit Gurdasani, Dave Palfrey, and Arpit Mittal. Demand-Weighted Completeness Prediction for a Knowledge Base. *NAACL*, 2018.

[15] Krzysztof Janowicz, Bo Yan, Blake Regalia, Rui Zhu, and Gengchen Mai. Debiasing knowledge graphs: Why Female Presidents are not like Female Popes. In *ISWC*, 2018.

[16] Maximilian Klein and Piotr Konieczny. Monitoring the Gender Gap with Wikidata Human Gender Indicators. *OpenSym*, 2016.

[17] Douglas B. Lenat and Edward A. Feigenbaum. On the thresholds of knowledge. *Artificial Intelligence*, 1991.

[18] Michael Luggen, Djellel Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux. Non-parametric Class Completeness Estimators for Collaborative Knowledge Graphs—The Case of Wikidata. In *ISWC*, 2019.

[19] Peter Mika and Jim Hendler. The Semantic Web Challenge 2008. *JWS*, 2008.

[20] Paramita Mirza, Simon Razniewski, Fariz Darari, and Gerhard Weikum. Cardinal Virtues: Extracting Relation Cardinalities from Text. *ACL*, 2017.

[21] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated MAchine reading COmprehension dataset. In *CoCo@NIPS*, 2016.

[22] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *International Conference on Scalable Information Systems*, 2006.

[23] Alan Patterson, Anant Narayanan, Anshu Jain, Jamie Taylor, Natasha Noy, and Yuqing Gao. Industry-scale Knowledge Graphs: Lessons and Challenges. *Communications of the ACM*, 2019.

[24] Heiko Paulheim. How much is a Triple? *ISWC*, 2018.

[25] Thomas Pellissier Tanon and Lucie-Aimée Kaffee. Property Label Stability in Wikidata: Evolution and Convergence of Schemas in Collaborative Knowledge Bases. *Wiki workshop at WWW*, 2018.

[26] Simon Razniewski. Optimizing update frequencies for decaying information. In *CIKM*, 2016.

[27] Simon Razniewski, Nitisha Jain, Paramita Mirza, and Gerhard Weikum. Coverage of information extraction from sentences and paragraphs. *EMNLP*, 2019.

[28] Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z Pan, Archit Sakhadeo, and Gerhard Weikum. Commonsense Properties from Query Logs and Question Answering Forums. *CIKM*, 2019.

[29] Fabian Suchanek. The Need to Move Beyond Triples. *Text2Story*, 2020.

[30] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: a core of semantic knowledge. *WWW*, 2007.

[31] Denny Vrandecic. Capturing Meaning: Toward an Abstract Wikipedia. In *ISWC Blue Sky ideas*, 2018.

[32] Denny Vrandecic. Collaborating on the sum of all knowledge across languages. *Wikipedia @ 20*, 2019.

[33] Denny Vrandečić and Markus Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 2014.

[34] Xiaolan Wang, Xin Luna Dong, Yang Li, and Alexandra Meliou. MIDAS: Finding the right web sources to fill knowledge gaps. In *ICDE*, 2019.

[35] Gerhard Weikum, Johannes Hoffart, and Fabian Suchanek. Ten Years of Knowledge Harvesting: Lessons and Challenges. *IEEE Data Engineering Bulletin*, 2016.

[36] Derry Tanti Wijaya, Ndapandula Nakashole, and Tom M Mitchell. "A Spousal Relation Begins with a Deletion of engage and Ends with an Addition of divorce": Learning State Changing Verbs from Wikipedia Revision History. *EMNLP*, 2015.

[37] Zhiyong Wu and others. PERQ: Predicting, explaining, and rectifying failed questions in KB-QA systems. In *WSDM*, 2020.

[9] https://worksheets.codalab.org/worksheets/0xba659fe363cb46e7a505c5b6a774dc8a