COOL-WD: A Completeness Tool for Wikidata

Fariz Darari[1,2](✉), Radityo Eko Prasojo[2], Simon Razniewski[2], and Werner Nutt[2]

[1] Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
[2] KRDB, Free University of Bozen-Bolzano, 39100, Italy
fariz@cs.ui.ac.id

Abstract. Wikidata as a collaborative Semantic Web data source has enjoyed increasing prominence, storing over 150 million facts about more than 26 million entities. Yet it is missing a way to assess its completeness. In this demo we present COOL-WD, a tool for supporting the completeness lifecycle of Wikidata, that is, the creation, view, update, and consumption of metadata about Wikidata completeness. COOL-WD is available at http://cool-wd.inf.unibz.it/ and has so far collected more than 10,000 completeness statements.

1 Introduction

The Semantic Web has quickly grown from merely 12 data sources in 2007 to more than 1000 in 2017.[1] It follows the open-world assumption: information contained is generally assumed to be incomplete. Consequently, knowledge base (KB) editors and consumers are oftentimes left clueless on which parts of information in the KB should be treated as complete. Indeed, many topics such as the children of Donald Trump, the crew of Apollo 11, or the states of Austria are found to be complete on Wikidata,[2] a collaborative KB with RDF support. Inspired from natural language completeness statements on the Web (as in, e.g., IMDb, OpenStreetMap, and Wikipedia) [2], it is therefore desirable to also describe the completeness of a KB via (machine-readable) completeness statements. The availability of such statements can add value to services that KBs are used for, such as entity recognition, question answering, or data journalism.

Completeness Statements. Completeness statements are expressed as (s, p) where s is a subject (or entity), and p is a predicate (or property). Intuitively, an RDF data source D having the completeness statement (s, p) means that D is complete for *all* p-values of s that exist in reality. For example, Wikidata is actually complete for all states of Austria, hence the statement (Q40,P150), where Q40 is the Wikidata identifier for Austria, and P150 for "contains administrative territorial entity", is applicable here.

[1] http://lod-cloud.net/
[2] https://www.wikidata.org/

Completeness Lifecycle. When talking about completeness statements of RDF data sources, we conceive that the statements go through a lifecycle consisting of four phases as shown in Fig. 1:

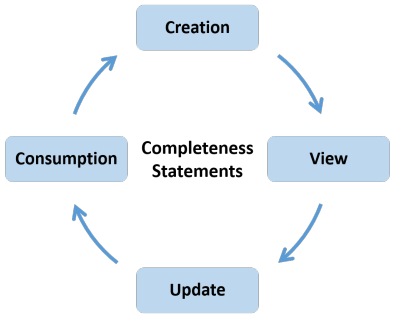

Fig. 1. Completeness Lifecycle

1. Creation: Completeness statements about RDF data sources are created, where provenance (e.g., author, reference, timestamp) of the statements can also be added.

2. View: Completeness statements (and their provenance), as well as parts of data captured by the statements, are available for viewing.

3. Update: Completeness statements can be updated (i.e., edited or deleted), if, for example, they are no longer valid or incorrectly given.

4. Consumption: Various consumption tasks such as query completeness checking and completeness analytics are performed.

The stages are collectively referred to as cycle because the consumption stage might reveal (in-)completeness problems of the data source, thus triggering the further creation of completeness statements.

Outcome. The intended outcome of the demo session is three-fold, that visitors shall become aware of: *(i)* how natural completeness information is for KBs (e.g., Wikidata); *(ii)* how the completeness lifecycle models a process of creating, viewing, updating, and consuming completeness information; and *(iii)* how COOL-WD realizes support for the completeness lifecycle of Wikidata.

While the present demonstration focuses on Wikidata, the need for expressing and storing completeness information is not limited to this KB. Our goal is to increase awareness of the completeness issue on the Semantic Web, and the need for considering it in the design of future methodologies, standards, and KBs.

Related Work. Data completeness relates to the breadth, depth, and scope of information in the data [4]. Techniques to measure completeness of RDF data sources, as surveyed by Zaveri et al. [5], are commonly done via comparison with the gold-standard data source. The surveyed techniques did not concern how to express that a source is of gold-standard completeness quality. Formal approaches that deal with the representation and reasoning of completeness statements have been proposed both for relational DBs [3] and RDF KBs [2]. Such approaches, however, lack a practical methodology of working with completeness statements. Recoin [1], a *relative* completeness tool for Wikidata, looks at the extent of information about Wikidata entities as a whole, for instance, whether or not information about Donald Trump is more abundant than about other, similar entities.

2 COOL-WD

COOL-WD is a completeness tool for Wikidata, enabling collaborative annotations of complete parts of Wikidata. COOL-WD maintains real time communication with Wikidata, and is implemented as client-server architecture. Two in-sync clients are available: the COOL-WD Web interface, and direct editing from `wikidata.org`. The direct editing script can be activated by adding the line `importScript( 'User:Fadirra/coolwd.js' );` to the Wikidata user's `common.js` file. The server side is responsible for controlling the application logic and storing completeness information. The UI is developed using GWT.[3] The server is Apache Tomcat-based, and the completeness DB relies on PostgreSQL.

COOL-WD is in active use by the Wikidata community and contains so far around 10,000 completeness statements, available for bulk download at `http://completeness.inf.unibz.it/rdf-export/`.

3 Demonstration Experience

> *On the Wikidata page of Austria*[4]*, the user is left clueless whether all the states of Austria are present. Moreover, it is unknown which of the states are already complete for their districts, hindering where to put focus on data collection about Austria.*

1. Creation. By comparing information in Wikidata with the official website of Austria,[5] users can observe that Wikidata is complete for the states. Thus the user can add the corresponding completeness statement, as well as the reference URL. Using the direct editing, the adding is as simple as clicking the respective property to the Wikidata page about Austria. The Wikidata username and timestamp are added accordingly.

2. View. As illustrated in Fig. 2 (a), the user can view completeness statements about Austria, as well as the provenance, by clicking the "(i)" icon. Greenhiglighted properties indicate completeness, while yellow indicates unknown. For instance, the user can see that the completeness of short names is unknown. We also provide a Linked Data API for COOL-WD completeness statements. For example, the dereferenceable URI for the statement (`Q40`,`P150`) as above is `http://cool-wd.inf.unibz.it/resource/statement-Q40-P150`. There, the user can view an RDF description of the completeness statement, modeled using our completeness vocabulary (`http://completeness.inf.unibz.it/sp-vocab`) and the W3C PROV vocabulary (`http://www.w3.org/ns/prov`).

3. Update. The user may notice a completeness statement for "diplomatic relation", which is not valid anymore, as Austria in the meantime has also established diplomatic relations with Kyrgyzstan[6]. The user can therefore remove the

[3] `http://www.gwtproject.org/`
[4] `https://www.wikidata.org/wiki/Q40`
[5] `https://www.parlament.gv.at/ENGL/PERK/BOE/`
[6] `https://www.embassypages.com/missions/embassy11893`

Fig. 2. (a) Completeness statement of Austria as shown using the client inside the Wikidata page; and (b) completeness information for the query of all states of Austria.

statement simply by clicking the green box. To update the URL of an existing statement, the user can edit the reference field after clicking "(i)".

4. Consumption. So far, COOL-WD supports three consumption tasks: data completeness tracking, completeness analytics, and query completeness diagnostics. The tracking shows the progress of completing the data of a single entity in terms of the fraction of the (known) non-functional properties that are complete. The completeness analytics summarizes in a tabular view the completeness of a class of entities. For example, for the class of Austrian states, the user might be interested in the completeness percentage of their districts and borders (i.e., how complete is the class for these properties?). The percentage is computed based on whether completeness statements exist or not for the properties of interest wrt. all entities of the class. Finally, query completeness diagnostics enables the completeness assessment of SPARQL basic graph pattern queries (as in Fig. 2 (b)). Additionally, statements used to guarantee query completeness can be shown.

Acknowledgements This work was partially funded by the Free University of Bolzano under the TaDaQua project, and Universitas Indonesia under the XYZ project.

References

1. A. Ahmeti, S. Razniewski, and A. Polleres. Assessing the completeness of entities in knowledge bases. In *ESWC Posters & Demos*, 2017.
2. F. Darari, W. Nutt, G. Pirrò, and S. Razniewski. Completeness statements about RDF data sources and their use for query answering. In *ISWC*, 2013.
3. S. Razniewski and W. Nutt. Completeness of queries over incomplete databases. *PVLDB*, 4(11):749–760, 2011.
4. R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *J. of Management Information Systems*, 12(4):5–33, 1996.
5. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for Linked Data: A survey. *Semantic Web*, 7(1):63–93, 2016.