# Question Answering Benchmarks for Wikidata

Dennis Diefenbach[1], Thomas Pellissier Tanon[2], Kamal Singh[1], Pierre Maret[1]

[1] Université de Lyon, CNRS UMR 5516 Laboratoire Hubert Curien
{dennis.diefenbach,kamal.singh,pierre.maret}@univ-st-etienne.fr
[2] Université de Lyon, ENS de Lyon, Inria, CNRS, Université Claude-Bernard Lyon 1, LIP.
thomas.tanon@ens-lyon.fr

**Abstract.** Wikidata is becoming an increasingly important knowledge base whose usage is spreading in the research community. However, most question answering systems evaluation datasets rely on Freebase or DBpedia. We present two new datasets in order to train and benchmark QA systems over Wikidata. The first is a translation of the popular SimpleQuestions dataset to Wikidata, the second is a dataset created by collecting user feedbacks.

**Keywords:** Wikidata, Question Answering Datasets, SimpleQuestions, WebQuestions, QALD, SimpleQuestionsWikidata, WDAquaCore0Questions

## 1 Introduction

Wikidata is becoming an increasingly important knowledge base. Its popularity has grown after the termination of Freebase [2] since August 31, 2016 and an effort has been made to migrate its content to Wikidata [9]. The information in Wikidata is also partially overlapping with the one of DBpedia. One of the goals of Wikidata is to generate infobox of Wikipedia using the information in the Wikidata Knowledge Base. Since this is the main source of information of DBpedia the overlap will increase over time. Alltogether this means that Wikidata is becoming an important sources for general knowledge in the Semantic Web.

On the other side the most popular benchmarks for QA systems namely WebQuestions, SimpleQuestions and QALD are mainly considering Freebase and DBpedia as datasets.

We therefore see a strong need in datasets for evaluating QA systems over Wikidata. In the following, we present two new datasets for QA over Wikidata. The first is a translation of the SimpleQuestion dataset from Freebase to Wikidata, the second is a dataset that was created exploiting user feedback mechanism in a QA system.

## 2 Related work

The most popular benchmarks for QA over Knowledge Bases are WebQuestions [1], SimpleQuestions [3] and QALD[3] [8][5][10][11][12]. Both WebQuestions and SimpleQuestions were designed for Freebase. WebQuestions contains 5810 questions. They can be answered using one reified statement with potentially some constraints like type constraints or temporal constraints[4]. SimpleQuestions contains 108.442 questions which can be answered using one triple pattern.

[3] http://www.sc.cit-ec.uni-bielefeld.de/qald/

[4] https://www.microsoft.com/en-us/download/details.aspx?id=52763

Another popular benchmark is QALD. The number of questions and the datasets used in the different QALD challenges are reported in Table 1. They generally can be answered using up to 3 triple patterns. Sometimes modifiers like COUNT and aggregation operators are needed. Note that only in the last QALD challenge a benchmark for QA over Wikidata was presented. It contains 150 questions.

Some other less known benchmarks on top of Freebase exists, like Free917 [4] that provides 917 questions annotated with lambda calculus forms.

Table 1: Overview of the QALD benchmarks launched so far.

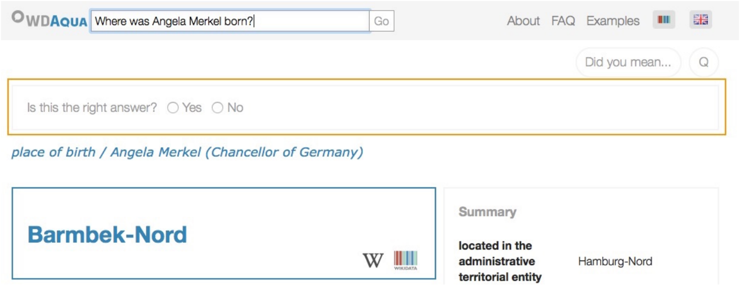
| Challenge | Task | Dataset | Questions | Languages |
|---|---|---|---|---|
| QALD-1 | 1 | DBpedia 3.6 | 50 train / 50 test | en |
| | 2 | MusicBrainz | 50 train / 50 test | en |
| QALD-2 | 1 | DBpedia 3.7 | 100 train / 100 test | en |
| | 2 | MusicBrainz | 100 train / 100 test | en |
| QALD-3 | 1 | DBpedia 3.8 | 100 train / 99 test | en, de, es, it, fr, nl |
| | 2 | MusicBrainz | 100 train / 99 test | en |
| QALD-4 | 1 | DBpedia 3.9 | 200 train / 50 test | en, de, es, it, fr, nl, ro |
| | 2 | SIDER, Diseasome, Drugbank | 25 train / 50 test | en |
| | 3 | DBpedia 3.9 with abstracts | 25 train / 10 test | en |
| QALD-5 | 1 | DBpedia 2014 | 300 train / 50 test | en, de, es, it, fr, nl, ro |
| | 2 | DBpedia 2014 with abstracts | 50 train / 10 test | en |
| QALD-6 | 1 | DBpedia 2015 | 350 train / 100 test | en, de, es, it, fr, nl, ro, fa |
| | 2 | DBpedia 2015 with abstracts | 50 train / 25 test | en |
| | 3 | LinkedSpending | 100 train / 50 test | en |
| QALD-7 | 1 | DBpedia 2016-04 | 214 train / 100 test | en, de, es, it, fr, nl, $\mathrm{hi}_I N$ |
| | 2 | DBpedia 2016-04 with abstracts | 100 train / 50 test | en |
| | 3 | DBpedia 2016-04 | Syntetic questions | en |
| | 4 | Wikidata | 100 train / 50 test | en |

## 3 SimpleQuestions to Wikidata

In this section, we describe how we ported the SimpleQuestions dataset originally designed for Freebase to Wikidata. The SimpleQuestions dataset [3] provides 108,442 questions, each annotated with a Freebase triple such that one of the acceptable answers to the question is the subject of the triple.

We mapped the Freebase triples to Wikidata using the same mapping process as [9]: the subject and objects of triples, that are Freebase topics are mapped to Wikidata items using automatically generated mappings and the properties are mapped using a handmade mapping. When there is no equivalent property in Wikidata, but the Freebase inverse property has an equivalent property `P`*XX* in Wikidata, we map the Freebase property to a "fake" Wikidata property `R`*XX* ("`R`" indicating reverse). Note that not every translated triple is required to exist in Wikidata. When migrating the data from Freebase to Wikidata part of the information was lost. We therefore have created two versions, the first containing questions that can be answered in Wikidata, the second containing all questions. The QA community has concentrated so far on benchmarking QA systems assuming that most of the questions in the benchmark are answerable. But having many questions that are not answerable allows to tackle a new challenge, i.e. let the QA system decide if it has the knowledge to answer the question or not.

This new dataset contains 49.202 questions (21.957 of which are answerable over Wikidata). One of the reason of the gap in size between the two datasets is that we have only mapped 404 Freebase properties to Wikidata even if the dataset contains 1.837 properties. But the top 50 properties in the Freebase dataset provide 61% of the triples and the top 100, 76%. It allowed to map 45% of the dataset with 108 properties.



Where was Angela Merkel born?

Is this the right answer? ○ Yes ○ No

*place of birth / Angela Merkel (Chancellor of Germany)*

**Barmbek-Nord**

Summary

| | |
|---|---|
| **located in the administrative territorial entity** | Hamburg-Nord |

Fig. 1: Snapshot of the Trill front-end. In orange the UI-component for feedback is highlighted.

The datasets is available under the Creative Commons Attribution 3.0 licence at `https://github.com/askplatypus/wikidata-simplequestions`. We offer it in the same format as the original SimpleQuestions dataset and in QALD format. We call it SimpleQuestionsWikidata.

## 4 Dataset using Logs and User feedback

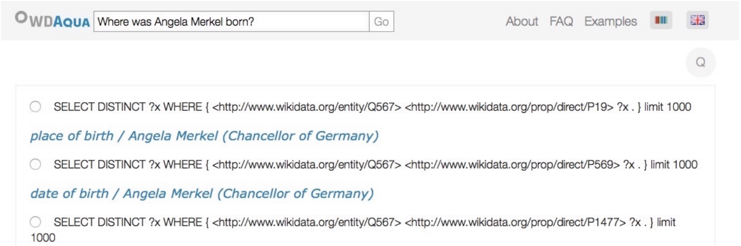
Creating large datasets for QA is a tedious and expensive task. For example, [1] report that they spent several thousands dollars for the creation of WebQuestions (containing 5810 questions) using Amazon Mechanical Turk. In the following, we show how it is possible to create a benchmark dataset for QA reducing the human effort and therefore also the cost.

The idea is to involve users of a QA system in the creation of the benchmark. In this concrete case we collected the feedback given by users using WDAqua-core0 [7], a QA system available under `www.wdaqua.eu/qa`. The web-service is online since June 2016 and received 12302 requests (5231 from them are unique). The QA service is exposed using Trill [6], a reusable front-end for QA systems. It contains a UI-component for feedback (see figure 1). The UI-component can be used mainly in two situations. In the first, end-users use the feedback component if they know that the answer is right or not. The second situation involves expert users that are able to understand SPARQL queries. If the answer is correct, the expert users can directly use the feedback component. If not they can check the top-k SPARQL queries generated by the QA system, as shown in Figure 2 and (if available) select the right one.

The collected data contains 689 questions. Note that the questions are asked by real users and are a mixture between keyword and full natural language questions. The questions can be answered with maximal 2 triple patterns. Considering how the dataset was created the benchmark can contain errors. The resulting benchmark is available in QALD format at `https://github.com/WDAqua/WDAquaCore0Questions` under the Creative Commons Attribution 3.0 licence. We call the generated dataset WDAquaCore0Questions.

## 5 Conclusion

We have presented two new datasets for the research community. The first dataset is the translation of a popular benchmark over Freebase to Wikidata, namely SimpleQuestions which contains 21.957 questions answerable over Wikidata. The second is a dataset containing 689 questions generated using user feedback. By offering these datasets we hope to move the QA community towards Wikidata, which



Where was Angela Merkel born?

○ SELECT DISTINCT ?x WHERE { <http://www.wikidata.org/entity/Q567> <http://www.wikidata.org/prop/direct/P19> ?x . } limit 1000

*place of birth / Angela Merkel (Chancellor of Germany)*

○ SELECT DISTINCT ?x WHERE { <http://www.wikidata.org/entity/Q567> <http://www.wikidata.org/prop/direct/P569> ?x . } limit 1000

*date of birth / Angela Merkel (Chancellor of Germany)*

○ SELECT DISTINCT ?x WHERE { <http://www.wikidata.org/entity/Q567> <http://www.wikidata.org/prop/direct/P1477> ?x . } limit 1000

Fig. 2: Snapshot of the Trill front-end. The UI component allowing one to select between the generated SPARQL queries is opened. By clicking on the corresponding SPARQL query the corresponding result set is computed and shown to the user. Afterwards the UI component for feedback can be used.

is becoming an increasingly important general Knowledge Base in the Semantic Web.

**Acknowledgments** Parts of this work received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 642795, project: Answering Questions using Web Data (WDAqua). It was also supported by the LABEX MILYON (ANR-10-LABX-0070) of Université de Lyon.

## References

1. Berant, J., Chou, A., Frostig, R., Liang, P.: Semantic Parsing on Freebase from Question-Answer Pairs. In: EMNLP (2013)
2. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. pp. 1247–1250. AcM (2008)
3. Bordes, A., Usunier, N., Chopra, S., Weston, J.: Large-scale simple question answering with memory networks. arXiv preprint arXiv:1506.02075 (2015)
4. Cai, Q., Yates, A.: Large-scale Semantic Parsing via Schema Matching and Lexicon Extension. In: ACL (1). Citeseer (2013)
5. Cimiano, P., Lopez, V., Unger, C., Cabrio, E., Ngomo, A.C.N., Walter, S.: Multilingual question answering over linked data (qald-3): Lab overview. Springer (2013)
6. Diefenbach, D., Amjad, S., Both, A., Singh, K., Maret, P.: Trill: A reusable front-end for qa systems. In: ESWC P&D (2017)
7. Diefenbach, D., Singh, K., Maret, P.: Wdaqua-core0: A question answering component for the research community. In: ESWC, 7th Open Challenge on Question Answering over Linked Data (QALD-7) (2017)
8. Lopez, V., Unger, C., Cimiano, P., Motta, E.: Evaluating question answering over linked data. Web Semantics: Science, Services and Agents on the World Wide Web (2013)
9. Pellissier Tanon, T., Vrandečić, D., Schaffert, S., Steiner, T., Pintscher, L.: From Freebase to Wikidata: The great migration. In: Proc. of WWW. pp. 1419–1428 (2016)
10. Unger, C., Forascu, C., Lopez, V., Ngomo, A.C.N., Cabrio, E., Cimiano, P., Walter, S.: Question answering over linked data (QALD-4). In: Working Notes for CLEF 2014 Conference (2014)
11. Unger, C., Forascu, C., Lopez, V., Ngomo, A.C.N., Cabrio, E., Cimiano, P., Walter, S.: Answering over Linked Data (QALD-5). In: Working Notes for CLEF 2015 Conference (2015)
12. Unger, C., Ngomo, A.C.N., Cabrio, E., Cimiano: 6th Open Challenge on Question Answering over Linked Data (QALD-6). In: The Semantic Web: ESWC 2016 Challenges. (2016)