# 7th Open Challenge on Question Answering over Linked Data (QALD-7)

Ricardo Usbeck[1], Axel-Cyrille Ngonga Ngomo[1], Bastian Haarmann[2], Anastasia Krithara[3], Michael Röder[1], and Giulio Napolitano[2]

Data Science Group, Paderborn University, Germany
usbeck|ngonga|roeder@informatik.uni-leipzig.de
Fraunhofer-Institute IAIS, Sankt Augustin, Germany
bastian.haarmann|giulio.napolitano@iais.fraunhofer.de
National Center for Scientific Research "Demokritos", Athens, Greece
akrithara@iit.demokritos.gr

## 1 Introduction

The past years have seen a growing amount of research on question answering (QA) over Semantic Web data, shaping an interaction paradigm that allows end users to profit from the expressive power of Semantic Web standards while, at the same time, hiding their complexity behind an intuitive and easy-to-use interface. On the other hand, the growing amount of data has led to a heterogeneous data landscape where QA systems struggle to keep up with the volume, variety and veracity of the underlying knowledge.

The Question Answering over Linked Data (QALD) challenge aims at providing an up-to-date benchmark for assessing and comparing state-of-the-art-systems that mediate between a user, expressing his or her information need in natural language, and RDF data. It thus targets all researchers and practitioners working on querying Linked Data, natural language processing for question answering, multilingual information retrieval and related topics. The main goal is to gain insights into the strengths and shortcomings of different approaches and into possible solutions for coping with the large, heterogeneous and distributed nature of Semantic Web data.

QALD[1] has a 6-year history of developing a benchmark that is increasingly being used as standard evaluation tool for question answering over Linked Data. Overviews of the past instantiations of the challenge are available from the CLEF Working Notes as well as ESWC proceedings:

- QALD-6: http://www.springer.com/us/book/9783319465647
- QALD-5: http://ceur-ws.org/Vol-1391/173-CR.pdf
- QALD-4: http://ceur-ws.org/Vol-1180/CLEF2014wn-QA-UngerEt2014.pdf
- QALD-3: https://pub.uni-bielefeld.de/download/2685575/2698020

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

Furthermore, through the QALD challenge, we (1) provide objective measures for how well current systems perform on real tasks of industrial relevance and (2) detect bottlenecks of existing systems in order to further develop them and make them more usable in practice. Since many of the topics relevant for QA over Linked Data lie at the core of ESWC (Multilinguality, Semantic Web, Human-Machine-Interfaces), we have run the 7th instantiation of QALD again at ESWC 2017. This year the challenge was supported by the EU project HOBBIT [1], which has already established a network of people from the Semantic Web as well as the Big Data community, both from the academia and industries. In addition, HOBBIT provided an open source holistic benchmarking platform for Big Linked Data, in which the challenge was run. Thanks to the HOBBIT project we were able to guarantee a controlled setting involving rigorous evaluations via its platform.[2]

*Similar Events.* To the best of our knowledge, there is no event with a comparable scope (Linked, Large-Scale, Hybrid Data) outside this series in the Semantic Web Community. However, there has thus been a number of challenges and campaigns attracting researchers as well as industry practitioners to QA. Since 1998, the TREC conference, especially the QA track [4], aims at providing domain-independent evaluations over large, unstructured corpora as well as Community-based QA. Next to that, the BioASQ series [2] challenges semantic indexing as well as QA systems on biomedical data and is currently at its fifth installment. Here, systems have to work on RDF as well as textual data to present matching triples as well as text snippets. The OKBQA challenge[3] is primarily an open QA platform powered by several Korean research institutes but they also released the NLQ datasets.

## 2 Tasks and Datasets

The key challenge for QA over Linked Data is to translate a user's information need into such a form that it can be evaluated using standard Semantic Web query processing and inferencing techniques. The main task of QALD therefore is the following:

> Given one or several RDF dataset(s) as well as additional knowledge sources and natural language questions or keywords, return the correct answers or a SPARQL query that retrieves these answers.

### Data format

All data for the tasks can be found in our project repository `https://github.com/ag-sc/QALD/tree/master/7/data`. We encouraged the use of QALD-JSON

[2] `https://project-hobbit.eu/challenges/qald2017/`
[3] `http://www.okbqa.org`

format[4] as communication format between the systems and the GERBIL QA respectively HOBBIT platform:

```
{ "id":"3",
  "answertype":"resource",
  "aggregation":false,
  "onlydbo":true,
  "hybrid":false,
  "question":[
    {
      "language":"en",
      "string":"Who was the wife of U.S. president
         Lincoln?",
      "keywords":"U.S. president, Lincoln, wife"
    },
    {
      "language":"nl",
      "string":"Wie was de vrouw van de Amerikaanse
         president Lincoln?",
      "keywords":"vrouw, president van America, Lincoln
         "
    }
  ],
  "query":{
    "sparql":"PREFIX dbo:<http://dbpedia.org/ontology/>
        PREFIX res:<http://dbpedia.org/resource/>
        SELECT DISTINCT ?uri
        WHERE {res:Abraham_Lincoln dbo:spouse ?uri.}"
  },
  "answers":[
    {
      "head":{
        "vars":[
          "uri"
        ]
      },
      "results":{
        "bindings":[
          {
            "uri":{
              "type":"uri",
              "value":"http://dbpedia.org/resource/
                 Mary_Todd_Lincoln"
            }
```

[4] https://github.com/AKSW/gerbil/wiki/Question-Answering and the results are formatted according to https://www.w3.org/TR/sparql11-results-json/

```
            }
        ]
    }}]}
```

In order to focus on specific aspects and challenges, we included the following four tasks.

**Task 1: Multilingual question answering over DBpedia** Given the diversity of languages used on the web, there is an increasing need to facilitate multilingual access to semantic data. The core task of QALD is thus to retrieve answers from an RDF data repository given an information need expressed in a variety of natural languages.

*Training data.* The underlying RDF dataset was DBpedia 2016-04. The training data consists of 215 questions compiled and curated from previous challenges. The questions are be available in eight different languages (English, Spanish, German, Italian, French, Dutch, Romanian and Farsi). Those questions are general, open-domain factual questions, for example:

(en) *Which book has the most pages?*
(de) *Welches Buch hat die meisten Seiten?*
(es) *¿Que libro tiene el mayor numero de paginas?*
(it) *Quale libro ha il maggior numero di pagine?*
(fr) *Quel livre a le plus de pages?*
(nl) *Welk boek heeft de meeste pagina's?*
(ro) *Ce carte are cele mai multe pagini?*

The questions vary with respect to their complexity, including questions with counts (e.g., *How many children does Eddie Murphy have?...*), superlatives (e.g., *Which museum in New York has the most visitors?*), comparatives (e.g., *Is Lake Baikal bigger than the Great Bear Lake?*), and temporal aggregators (e.g., *How many companies were founded in the same year as Google?*). Each question is annotated with a manually specified SPARQL query and answers. In the above case, the SPARQL query looks as follows:

```
SELECT DISTINCT ?uri
WHERE {
    ?uri a <http://dbpedia.org/ontology/Book> .
    ?uri <http://dbpedia.org/ontology/numberOfPages> ?n .
}
ORDER BY DESC(?n)
OFFSET 0 LIMIT 1
```

And the answer is `<http://dbpedia.org/resource/The_Tolkien_Reader>`.

*Test Data.* The test dataset consists of 50 similar manually created questions. However, this year we decided to increase the complexity of the test data and add several other question types including questions according to RDF types

(e.g., *What is backgammon?...*) or questions demanding mathematical operations (e.g., *What is the radius of the earth?...*). They are compiled from existing, real-world question and query logs, in order to provide unbiased questions expressing real-world information needs. The questions were manually curated to ensure a high quality standard.

### Task 2: Hybrid question answering

A large amount of information is still available as unstructured text only, both on the web and in the form of labels and abstracts in Linked Data sources. Therefore, approaches are needed that can not only deal with the specific character of structured data but also with finding information in other sources, processing both structured and unstructured information, and combining such gathered information into a single answer. Therefore, QALD-7 included a task on hybrid question answering, forcing systems to retrieve answers for questions that required the integration of data both from RDF and from textual sources.

*Training data.* The training data is build using DBpedia 2016-04 as the RDF knowledge base, together with the English Wikipedia as the textual data source. As training data, we included 105 questions in English from past challenges (partly based on questions used in the INEX Linked Data track[5]). The questions are annotated with answers as well as a pseudoquery that indicates what information can be obtained from RDF data and what from free text. The pseudoquery is like an RDF query but may contain free text as subject, property or object of a triple. An example is the question *Who is the front man of the band that wrote Coffee & TV?*, with the following corresponding pseudoquery:

```
SELECT DISTINCT ?uri
WHERE {
    <http://dbpedia.org/resource/Coffee_&_TV>
    <http://dbpedia.org/ontology/musicalArtist> ?x .
    ?x <http://dbpedia.org/ontology/bandMember> ?uri .
    ?uri text:"is" text:"frontman" .
}
```

The manually specified answer is `<http://dbpedia.org/resource/Damon_Albarn>`.

*Test data.* As test questions, we generated 50 similar questions, all manually created and checked by at least two data experts. The main goal in devising those questions was not to take into account the vast amount of data available and the problems arising from noisy, duplicate and conflicting information. Rather, we aimed at enabling a controlled and fair evaluation, considering that hybrid question answering is still a very young line of research.

[5] `http://inex.mmci.uni-saarland.de/tracks/dc/index.html`

**Task 3: Large-Scale Question answering over RDF**

A new task was introduced this year, with focus on large-scale question sets. The aim was to assess approaches able to scale up to a big data volume, handle a vast amount of questions and speed up the question answering process by parallelization, such that the highest possible number of questions can be answered as accurately as possible in the shortest possible time. Again, the data for this task is based on the DBpedia 2016-04 RDF knowledge base.

*Data creation.* The training set consists of 100 questions compiled from the HOBBIT project. The test set 0f 2M questions is generated by an algorithm deriving new questions from the training set by varying both the query desire and the form of the natural language expression. Questions were annotated with SPARQL queries and answers and, in the test scenario, they were sent every minute to the competing systems in packets of increasing size, with *n+1* questions asked at minute *n*. Participating systems were evaluated with respect to both number of correct answers and time needed.

**Task 4: Question answering over Wikidata** This task, also new for the 7th edition of the QALD challenges, provided a benchmark focusing on the ability of systems to adapt to new data sources. Questions originally formulated for DBpedia require an answer using Wikidata, so that systems have to deal with a different data representation structure. The task is meant to support the evaluation of how generic the approach of a given system is and how easy it is to adapt to a new data source.

*Data creation.* The training question set consisted of 100 questions selected from Task 1 of the QALD-6 challenge. We formulated the queries to answer these questions from Wikidata and generated the gold standard answers using them on the Wikidata dump from 09-01-2017. As test data, 50 additional questions were used from the QALD-6 challenge.

| **Task** | **Train** | **test** |
|---|---|---|
| 1. Multilingual | 215 | 50 |
| 2. Hybrid | 105 | 50 |
| 3. Large-scale | 100 | 2M |
| 4. Wikidata | 100 | 50 |

**Table 1.** Number of questions per task in the training and test sets.

## 3 Evaluation

The QALD challenge provides an automatic evaluation tool (GERBIL QA [3] integrated into the HOBBIT platform)[6][7] that is open source and available for everyone to re-use. The GERBIL QA platform is accessible online, so that participants can simply upload the answers produced by their system or even check their system via a webservice. Each experiment has a citable, time-stable and archivable URI which is both human- and machine-readable. However, participating systems had to provide a Docker container[8][9] to participate in the final challenge which communicated with the HOBBIT platform.

The QA systems were evaluated with respect to precision and recall. For each question $q$, precision and recall are computed as follows:

$$\text{recall}(q) = \frac{\text{number of correct system answers for } q}{\text{number of gold standard answers for } q}$$

$$\text{precision}(q) = \frac{\text{number of correct system answers for } q}{\text{number of system answers for } q}$$

The evaluation computed the macro and micro F-measure of a system over all test questions. That is, for micro F-measure we summed up all true and false positives and negative up and calculated in the end the precision, recall and F-measure while for the macro measures we calculated precision, recall and F-measure per question and averaged the values in the end.

In this challenge, we left out the computation of measures over only those questions that the system did provide an answer for. That is, question without answer would have been ignored instead of resulting in a zero F-measure. Thus, it was not possible to make an evaluation which would have allowed to take into account a system's ability to identify questions that it cannot answer. For task 3, specifically, the evaluation takes into account not only the accuracy measures for the answered questions but also the scalability measures in terms of number of processed queries and time needed for answer retrieval.

## 4 Participating systems

Three teams participated in the QALD-7 challenge, with three teams addressing the multilingual task (two for English and one French) and two addressing the QA over Wikidata task. Note, that the description of the papers can be found in the challenge proceedings of the 2017 ESWC satellite proceedings.

**WDAqua** is a rule-based system using a combinatorial approach to generate SPARQL queries from natural language questions, leveraging the semantics

[6] `http://gerbil-qa.aksw.org/gerbil/`
[7] `http://project-hobbit.eu/`
[8] `https://project-hobbit.eu/challenges/qald2017/#Technical_requirements`
[9] `https://github.com/hobbit-project/platform/wiki/Participate-in-a-challenge`

encoded in the underlying knowledge base. It can answer questions on both DBpedia (supporting English) and Wikidata (supporting English, French, German and Italian). The system, which does not require training, participated in Task 1 and 4 of the challenge.

**AMAL** has been developed for QA in French. Firstly, the question type (e.g. *Boolean* or *Entity*) is classified by pattern matching. This induces the rerouting to the relevant question type solver where entities and properties are extracted: the former by syntactic parsing and subsequent linking to DBpedia entities; the latter by removing the found entity and searching for corresponding properties in DBpedia, possibly with the help of with Wikipage disambiguation links. SPARQL predicate identification is supported by a manually curated lexicon of common DBpedia properties, each linked to one or more possible French expressions. The system can only answer simple questions (concerning a single entity or a single property of an entity) and participated in Task 1.

**Sorokin and Gurevych** participated in Task 4 of the challenge. They provided a system producing the semantic representation of a natural language question, which is then deterministically converted into SPARQL. After minimal pre-processing, including POS tagging and entity linking, an end-to-end neural architecture employs a CNN neural scorer to choose among multiple semantic representations of the question. First, the semantic representations are generated by expansion on the knowledge base, guided by the entity found in the question and by all possible relations and constraints as present in the KB for the entity. Then, each question and candidate representations are vectorialised, with the CNN producing comparison scores based on cosine similarity, leading to the final choice.

**ganswer2** [5] has participated outside the actual challenge this year as a system without a paper submission in Task 1. Zou et al. use a graph-based approach to generate a semantic query graph which reduced the transformation of natural language to SPARQL to a subgraph matching problem.

## 5 Results

### Task 1: Multilingual question answering over DBpedia

Task 1 was run for the seventh time in 2017. Three participating teams submitted their systems via the HOBBIT or GERBIL QA platform. Please note that AMAL submitted their results as files, due to constraints of the system which resulted in using GERBIL QA as a platform. Also note that WDAQUA macro values are taken from the system authors challenge submission, as we could not use the platform at the time of this publication. Furthermore, we used GERBIL QA for the training data evaluation as HOBBIT was only targeted for the evaluation of the actual challenge (blind test) data.

The experimental data for task 1 over training data can be found in the following:

– ganswer (en): `http://gerbil-qa.aksw.org/gerbil/experiment?id=201706300001`,

– AMAL (fr): http://gerbil-qa.aksw.org/gerbil/experiment?id=201706300002.

The experimental data for task 1 over the test data can be found in the following:

– ganswer (en): http://master.project-hobbit.eu/#/experiments/details?id=1498647986590,
– WDAqua (en): http://master.project-hobbit.eu/#/experiments/details?id=1498647742687,
– AMAL (fr): http://gerbil-qa.aksw.org/gerbil/experiment?id=201706300011.

By providing human- and machine-readable experimental URIs, we provide deeper insights and repeatable experiment setups.

Note also that the numbers reported here may differ from the publications of the participants, as these figures were not available at the time of participant paper submission.

| **Test** | **WDAqua** | **ganswer2** | **AMAL** |
|---|---|---|---|
| **Language** | **en** | **en** | **fr** |
| Error count | | 3 | |
| Micro Precision | 0.080 | 0.322 | 0.998 |
| Micro Recall | 0.006 | 0.127 | 0.989 |
| Micro F1-measure | 0.012 | 0.182 | 0.993 |
| Macro Precision | 0.162 | 0.487 | 0.720 |
| Macro Recall | 0.160 | 0.498 | 0.720 |
| Macro F1-measure | 0.143 | 0.469 | 0.720 |
| **Train** | **WDAqua** | **ganswer2** | **AMAL** |
| **Language** | **en** | **en** | **fr** |
| Error count | | | |
| Micro Precision | - | 0.113 | 0.971 |
| Micro Recall | - | 0.561 | 0.697 |
| Micro F1-measure | - | 0.189 | 0.811 |
| Macro Precision | 0.490 | 0.557 | 0.750 |
| Macro Recall | 0.540 | 0.592 | 0.751 |
| Macro F1-measure | 0.510 | 0.556 | 0.751 |

**Table 2.** Overview over QALD-7 task 1.

### Task 4: Question answering over Wikidata

Task 4 was run this year at QALD-7 for the first time and announced at short notice. Thus, it only attracted two teams. However, both teams performed well

on both the train and the test datasets. For the first time in QALD, Sorokin and Gurevych also used a neural network to answer questions over Wikidata. As can be seen from the numbers in table 5, both systems have a higher macro F-measure than micro F-measure. The task 4 data contains questions with long answer lists and if a system fails to answer such querys this has a huge impact on its micro recall and thus on its micro F-measure.

| **Test dataset** | **WDAqua** | **Sorokin and Gurevych** |
|---|---|---|
| Micro Precision | 0.392 | 0.428 |
| Micro Recall | 0.082 | 0.030 |
| Micro F1-measure | **0.136** | 0.057 |
| Macro Precision | 0.739 | 0.661 |
| Macro Recall | 0.606 | 0.430 |
| Macro F1-measure | **0.552** | 0.427 |
| **Train dataset** | **WDAqua** | **Sorokin and Gurevych** |
| Micro Precision | 0.172 | 0.295 |
| Micro Recall | 0.112 | 0.070 |
| Micro F1-measure | **0.136** | 0.113 |
| Macro Precision | 0.759 | 0.774 |
| Macro Recall | 0.710 | 0.756 |
| Macro F1-measure | 0.636 | **0.645** |

**Table 3.** Overview over QALD-7 task 4. The experimental data for task 4 executed with the HOBBIT platform can be found here `http://master.project-hobbit.eu/#/experiments/details?id=1498647794373`, `http://master.project-hobbit.eu/#/experiments/details?id=1498647917506`, `http://master.project-hobbit.eu/#/experiments/details?id=1498647883035` and `http://master.project-hobbit.eu/#/experiments/details?id=1498647941734`.

## 6 Summary

The seventh Question Answering over Linked Data challenge introduced two new tasks (scalable QA and QA over Wikidata) and repeated two of the successful past tasks. For the first time, the participating systems offered webservices as a prerequisite to participate in the challenge which will support comparable research in the future. In this challenge, we also changed the underlying evaluation platform to account for the need for comparable experiments via webservices and new technologies such as docker as compared to former XML/JSON file submissions. This increased the entranced barrier for participating teams but ensures a long term comparability of the system performance and a fair and open challenge.

In the future, we will further simplify the participation process and offer leader boards prior to the actual challenge in order to allow participants to

already see their performance. After feedback from the authors, we will add new key performance indicators to also account for the capability of a system to know which questions it cannot answer and take confidence scores for answers into account. Overall, we hope that the HOBBIT platform can serve as a long term challenge support to increase comparable and repeatable question answering research.

**Acknowledgments.** This work was supported by the Eurostars projects DIESEL (E!9367) and QAMEL (E!9725) as well as the European Union's H2020 research and innovation action HOBBIT under the Grant Agreement number 688227. We also want to thank Christina Unger and Sebastian Walter for supporting this challenge.

## References

1. Axel-Cyrille Ngonga Ngomo, Alejandra García-Rojas, and Irini Fundulaki. HOBBIT: holistic benchmarking of big linked data. *ERCIM News*, 2016(105), 2016.
2. George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel Ngonga, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138, 2015.
3. Ricardo Usbeck, Röder Michael, Christina Unger, Michael Hoffmann, Christian Demmler, Jonathan Huthmann, and Axel-Cyrille Ngonga Ngomo. Benchmarking question answering systems. Technical report, Leipzig University, 2016.
4. Ellen M Voorhees et al. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82, 1999.
5. Lei Zou, Ruizhe Huang, Haixun Wang, Jeffrey Xu Yu, Wenqiang He, and Dongyan Zhao. Natural language question answering over rdf: a graph data driven approach. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 313–324. ACM, 2014.