Question Answering Over Linked Data

QALD-2 Open Challenge

Document Version: March 26, 2012

The QALD-2 open challenge is the second instalment of the *Question Answering over Linked Data* benchmark and is organized as part of the workshop *Interacting with Linked Data* (ILD) at the Extended Semantic Web Conference 2012 in Heraklion, Greece. The challenge is aimed at any kind of question answering system that mediates between a user, expressing his or her information need in natural language, and semantic data. The goal is to evaluate and compare participating systems. To this end, two datasets are provided – DBpedia 3.7 and an RDF export of MusicBrainz – together with a set of test questions. Beforehand, a training set of 100 natural language questions per dataset is made publicly available, annotated with corresponding SPARQL queries and correct answers. All relevant information for accessing the datasets and questions as well as for participating in the challenge are given in this document.

1 Motivation and Goal

In contrast to the Information Retrieval community, where evaluation using standardized techniques, such as those used for the annual TREC competitions, has been common for decades, the Semantic Web community (characterized by the diversity of semantic technologies) has not yet adopted standard evaluation benchmarks for semantic question answering systems that focus on the ability to solve open-ended real life problems over real-world datasets. Aiming to make progress in this area and develop the datasets needed to formally judge the quality of ontology-based question answering approaches at a large scale, the QALD challenge provides tests consisting of a variety of questions of different complexity, designed to represent questions that real end users would ask.

The questions are provided along with keywords, in order to encourage both natural language and keyword-based approaches to participate. As some of the questions are indeed very challenging, feel free to work only on a subset, e.g., leaving out questions relying on Yago hierarchies and the like. We strongly encourage you to report on results even if precision and recall are relatively low, as the goal of the challenge is not to handle all complexities present in the queries, but rather to get a picture of the strengths, capabilities and current shortcomings of question answering systems, as well as to gain insight into how we can develop question answering approaches that deal with the fact that i) the amount of RDF data available on the Web is huge, ii) that this data is distributed and iii) that it is heterogeneous with respect to the vocabularies and schemas used.

Although the competition is tailored towards question answering systems based on natural language, we also encourage other relevant systems and methods, such as dynamic ontology matching, schema integration, word sense disambiguation, fusion and ranking technologies that can benefit from the QALD evaluation datasets to report their results.

2 Relevant information in a nutshell

Coordinators: Christina Unger, Philipp Cimiano, Vanessa Lopez, Enrico Motta, Paul Buitelaar, Richard Cyganiak

Workshop Website: `http://www.sc.cit-ec.uni-bielefeld.de/ild/`

Datasets:

- DBpedia 3.7
  `http://downloads.dbpedia.org/3.7/en/`
  `http://downloads.dbpedia.org/3.7/links/`
- MusicBrainz
  `http://greententacle.techfak.uni-bielefeld.de/~cunger/qald2/musicbrainz.tar.gz` (226.8 MB)

SPARQL end point (for both datasets):
`http://greententacle.techfak.uni-bielefeld.de:5171/sparql`

Training questions:
`http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/2/`

- `dbpedia-train.xml` and `dbpedia-train-answers.xml`
- `musicbrainz-train.xml` and `musicbrainz-train-answers.xml`

Test questions:
`http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/2/`

- `dbpedia-test-questions.xml`
- `musicbrainz-test-questions.xml`

Participant's challenge:
`http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/2/participants-challenge.xml`

Submission of results and evaluation is done by means of an online form:
`http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=evaltool&q=2`

Results for training questions can be uploaded at any time; results for test questions can be uploaded from March 25 to April 1.

Contact:
Updates on the open challenge will be published on the ILD mailing list:

`https://lists.techfak.uni-bielefeld.de/cit-ec/mailman/listinfo/ild`

In case of question, problems and comments, please contact Christina Unger: `cunger@cit-ec.uni-bielefeld.de`

3 Datasets

The open challenge uses two RDF datasets: DBpedia 3.7 (http://dbpedia.org) and MusicBrainz (musicbrainz.org). In order to work with the datasets, you can either download them or use the provided SPARQL end point.

3.1 DBpedia 3.7

DBpedia is a community effort to extract structured information from Wikipedia and to make this information available as RDF data. The RDF dataset provided for the challenge is the official DBpedia 3.7 dataset for English, including links, most importantly to YAGO categories and MusicBrainz. This dataset comprises all files contained in the following two directories:

- http://downloads.dbpedia.org/3.6/en/
- http://downloads.dbpedia.org/3.6/links/

The only exception are a few triples that contained overlong URIs and therefore could not be loaded into our Virtuoso store. If you want to see which triples are missing, you can download them from the following location:
http://greententacle.techfak.uni-bielefeld.de/∼cunger/download/dbpedia_toolong.zip

For more information on the DBpedia dataset, please refer to:

http://wiki.dbpedia.org

For detailed information on the YAGO class hierarchy, please see:

http://www.mpi-inf.mpg.de/yago-naga/yago/

Namespaces that are used in the provided training and test queries are the following:

```
 dbo: <http://dbpedia.org/ontology/>
 dbp: <http://dbpedia.org/property/>
 res: <http://dbpedia.org/resource/>
yago: <http://dbpedia.org/class/yago/>
foaf: <http://xmlns.com/foaf/0.1/>
 xsd: <http://www.w3.org/2001/XMLSchema#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
 rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

3.2 MusicBrainz

MusicBrainz is a collaborative effort to create an open content music database. The dataset provided for the challenge is an RDF export containing all classes

(artists, albums and tracks) and the most important properties of the MusicBrainz database. A package containing all RDF data[1] can be downloaded from the following location:
http://greententacle.techfak.uni-bielefeld.de/~cunger/qald2/musicbrainz.tar.gz (226.8 MB)

The RDF export builds no longer on the MusicBrainz ontology (as it did last year) but on the Music Ontology. The following namespaces are used in the provided trainingnd test queries:

```
   mo: <http://purl.org/ontology/mo/>
  bio: <http://purl.org/vocab/bio/0.1/>
  rel: <http://purl.org/vocab/relationship/>
event: <http://purl.org/NET/c4dm/event.owl#>
   tl: <http://purl.org/NET/c4dm/timeline.owl#>
 foaf: <http://xmlns.com/foaf/0.1/>
   dc: <http://purl.org/dc/elements/1.1/>
  xsd: <http://www.w3.org/2001/XMLSchema#>
  rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

A specification of the Music Ontology can be found at http://musicontology.com. Examples for how to model data w.r.t. this specification are given in the Music Ontology Wiki: http://wiki.musicontology.com/index.php/Examples. In the following, we briefly describe the most important classes and relations relevant for the challenge.

There are three major classes:

- `mo:MusicArtist` and its subtypes `mo:MusicGroup` for bands and orchestra, and `mo:SoloMusicArtist` for persons (independent of whether they are solo artists or members of a group)
- `mo:Record`
- `mo:Track`

Artists have a birth and death date modelled by means of the BIO vocabulary[2]. For example, the following SPARQL query extracts the birth and death date of John Lennon:

```
SELECT ?birthdate deathdate WHERE {
   ?artist foaf:name 'John Lennon'.
   ?artist bio:event ?event1 .
   ?event1 rdf:type bio:Birth .
   ?event1 bio:date ?birthdate .
   ?artist bio:event ?event2 .
   ?event2 rdf:type bio:Death .
   ?event2 bio:date ?deathdate .
}
```

[1] It contains only a subset of all track information, due to performance problems.

[2] http://vocab.org/bio/0.1/

In exactly the same way, the corresponding dates for music groups are formulated (where birth date can be read as founding date and death date as the date the group broke up).

Artists are related among each other through relations like `rel:spouseOf`, `rel:parentOf`, `siblingOf` and `rel:collaboratesWith` from the RELATIONSHIP vocabulary[3].

Membership in a group is expressed in two ways: by means of the simple relation `mo:member_of`, indicating that someone is or was member of a group, and by means of the Event and Timeline Ontology[4]. Using the former, the following triple expresses that John Lennon is or was a member of The Beatles:

```
<http://zitgist.com/music/artist/4d5447d7-c61c-4120-ba1b-d7f471d385b9>
mo:member_of
<http://zitgist.com/music/artist/b10bbbfc-cf9e-42e0-be17-e2c3e1d2600d>
```

Or, in a more human-readable way:

```
?artist foaf:name 'John Lennon' .
?artist mo:member_of ?band .
?band foaf:name 'The Beatles' .
```

In order to also express time information, the more complex Event and Timeline Ontology representation has to be used. For example, the following triples express that Pete Best was a member of The Beatles from August 12, 1960 until August 16, 1962.

```
?artist foaf:name 'Pete Best' .
?event rdf:type mo:membership .
?event event:agent ?artist .
?event mo:group ?band .
?band foaf:name 'The Beatles' .
?event event:time ?time .
?time tl:start '1960-08-12'^^xsd:date .
?time tl:end '1962-08-16'^^xsd:date .
```

Records are related to their creator through the property `foaf:maker`, and through `mo:releaseType` to the type of record (`mo:album`, `mo:single`, `mo:ep`, `mo:soundtrack`, `mo:live`, `mo:compilation`, `mo:remix`, `mo:interview`, and `mo:audiobook`). For example, the following SPARQL query extracts all live albums by Slayer:

```
SELECT ?album WHERE {
   ?album mo:release_type mo:live .
   ?album foaf:maker ?artist .
   ?artist foaf:name 'Slayer'.
}
```

[3] http://vocab.org/relationship/

[4] http://purl.org/NET/c4dm/event.owl and http://purl.org/NET/c4dm/timeline.owl.

The dataset also contains relations between records and artists, specifying their role during the record creation, for example mo:performer, mo:singer, mo:composer, mo:producer, and mo:lyricist.

Tracks are also related to their creator through the property foaf:maker, to their duration through tl:duration, and through mo:trackNum to their position in the track list of a record. For example, the following SPARQL query extracts the title of the first track of Abbey Road:

```
SELECT ?title WHERE {
   ?album dc:title 'Abbey Road' .
   ?album mo:track ?track .
   ?track mo:trackNum '1' .
   ?track dc:title ?title .
}
```

3.3 SPARQL end point

The SPARQL end point for both datasets is the following:

http://greententacle.techfak.uni-bielefeld.de:5171/sparql

Evaluation will take place with respect to this SPARQL end point (and not the official DBpedia end point, for example), in order to ensure invariable and therefore comparable results.

4 Evaluation

The task of the challenge is to extract correct answers for natural language questions or corresponding keywords from a given RDF repository. Participating systems will be evaluated with respect to precision and recall. Moreover, participants are encouraged to report performance, i.e. the average time their system takes to answer a query, if they submit a paper.

4.1 Training questions

In order to get acquainted with the datasets and possible questions, a set of 100 training questions for each dataset can be downloaded at the following locations: http://greententacle.techfak.uni-bielefeld.de/∼cunger/qald/2/

- DBpedia

  dbpedia-train.xml (without answers)
  dbpedia-train-answers.xml (with answers)

- MusicBrainz

musicbrainz-train.xml (without answers)
musicbrainz-train-answers.xml (with answers)

All training questions are annotated with keywords, corresponding SPARQL queries and, if indicated, answers retrieved from the provided SPARQL end point. Annotations are provided in the following XML format. The overall document is enclosed by a tag that specifies an ID for the dataset indicating the domain and whether it is train or test (i.e. `dbpedia-train`, `dbpedia-test`, `musicbrainz-train`, `musicbrainz-test`).

```
<dataset id="dbpedia-train">
  <question id="1"> ... </question>
  ...
  <question id="100"> ... </question>
</dataset>
```

Each of the questions specifies an ID for the question (don't worry if they are not ordered) together with a range of other attributes explained below, the natural language string of the question, keywords, a corresponding SPARQL query, as well as the answers this query returns. Here is an example:

```
<question id="36" answertype="resource" aggregation="false" onlydbo="false">
   <string>Through which countries does the Yenisei river flow?</string>
   <keywords>Yenisei river, flow through, country</keywords>
   <query>
    PREFIX res:  <http://dbpedia.org/resource/>
    PREFIX dbp:  <http://dbpedia.org/property/>
    SELECT DISTINCT ?uri ?string WHERE {
       res:Yenisei_River dbp:country ?uri .
       OPTIONAL { ?uri rdfs:label ?string . FILTER (lang(?string) = "en") }
    }
    </query>
   <answers>
     <answer>
      <uri>http://dbpedia.org/resource/Mongolia</uri>
      <string>Mongolia</string>
     </answer>
     <answer>
      <uri>http://dbpedia.org/resource/Russia</uri>
      <string>Russia</string>
     </answer>
   </answers>
</question>
```

The following attributes are specified for each question along with its ID:

- `answertype` gives the answer type, which can be one the following:
  - `resource`: One or many resources, for which both the URI as well as its English label (if it exists) is provided.

  - string: A string value such as Valentina Tereshkova.
  - number: A numerical value such as 47r 1.8.
  - date: A date provided in the format YYYY-MM-DD, for example 1983-11-02. This format is also required when you submit results containing a date as answer.
  - boolean: Either true or false.

  Answer of these types are required to be enclosed by the corresponding tag, i.e. <number>47</number>, <string>Valentina Tereshkova</string> and <boolean>true</boolean> (except for resources, for which a URI and/or a string should be provided, see the cave example above).

- aggregation indicates whether any operations beyond triple pattern matching are required to answer the question (e.g., counting, filters, ordering, etc.).
- onlydbo is given only for DBpedia questions and reports whether the query relies solely on concepts from the DBpedia ontology.

As an additional challenge, a few of the training and test questions are out of scope, i.e. they cannot be answered with respect to the dataset. The query is specified as OUT OF SCOPE and the answer set is empty. Here is an example from the DBpedia training question set:

```
<question id="94" answertype="number" aggregation="false" onlydbo="false">
   <string>Which budget did Zdenek Sverak’s first movie have?</string>
   <keywords>Zdenek Sverak, budget, first movie</keywords>
   <query>
   OUT OF SCOPE
   </query>
   <answers />
</question>
```

For evaluation, your system should in these cases specify OUT OF SCOPE as query and/or an empty answer set, just like in the example.

4.2 Submitting results

All submissions are required to comply with the XML format specified above. For all questions, the dataset ID and question IDs are obligatory. Beyond that, you are free to specify either a SPARQL query or the answers (or both), depending on which of them your system returns. You are also allowed to change the natural language question or keywords (insert quotes, reformulate, use some controlled language format, and the like). If you do so, please document these changes, i.e. replace the provided question string or keywords by the input you used.

Submissions will be accepted and evaluated by means of the following online form: http://greententacle.techfak.uni-bielefeld.de/∼cunger/qald2/evaluation/

4.3 Evaluation

For each of the questions, your specified answers, or the answers your specified SPARQL query retrieves, will be compared to the answers provided by the gold standard XML document. For resources, you are free to either provide their URI, their English label or name (if it exists), or both. All options count as correct answers, as long as the answer list contains all and only the correct resources.

The evaluation tool computes precision, recall and F-measure for every question:[5]

$$\begin{aligned}\text{Recall} &= \frac{\text{number of correct system answers}}{\text{number of gold standard answers}}\\ \text{Precision} &= \frac{\text{number of correct system answers}}{\text{number of system answers}}\\ \text{F-measure} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\end{aligned}$$

The tool then also computes the overall precision and recall taking the average mean of all single precision and recall values, as well as the overall F-measure.

All these results are printed in a simple HTML output; additionally you get a list of all question that your tool failed to capture correctly.

You are allowed to submit results as often as you wish.

4.4 Test phase

During test phase, a set of different questions for each dataset without annotations are provided at the following locations:

`http://greententacle.techfak.uni-bielefeld.de/∼cunger/qald/2/`

- `dbpedia-test-questions.xml`
- `musicbrainz-test-questions.xml`

Results can be submitted from March 25 to April 1, via the same online form used during training phase (note the drop down box that will allow you to specify *test* instead of *training*):

`http://greententacle.techfak.uni-bielefeld.de/∼cunger/qald/index.php?x=evaltool&q=2`

The only difference is that evaluation results are not displayed. You can upload results as often as you like (e.g., trying different configurations of your system); in this case the result file with best precision and recall will count.

[5] In the case of out-of-scope questions, an empty answer set counts as precision and recall 1, while a non-empty answer set counts as precision and recall 0.

4.5 Resources

You are free to use all resources (e.g., WordNet, GeoNames, dictionary tools, and so on).

5 Participant's challenge

Are there questions that your tool is very good at but that might prove difficult for others? Are there questions that are very interesting but are not among the training questions? Then send in these questions and challenge others!

In order to make a start, we provide a few questions that cannot be answered over DBpedia or MusicBrainz alone, but require the combination of both datasets. You can access them at the following location:
http://greententacle.techfak.uni-bielefeld.de/∼cunger/qald2/participants-challenge.xml

If there are any questions you would like to contribute, please send an email with the question (and, ideally, also a corresponding SPARQL query) to Christina Unger: cunger@cit-ec.uni-bielefeld.de. The questions will then be added to the document and published on the ILD mailing list.

6 Contact and trouble shooting

If you have any questions or comments, including worries about the training and test questions, trouble with the datasets, the SPARQL end point, or the online submission and evaluation form, please contact Christina Unger: cunger@cit-ec.uni-bielefeld.de.