The LSQ Dataset: Querying for Queries

Muhammad Saleem[1], Intizar Ali[2], Aidan Hogan[3],
Qaiser Mehmood[2], and Axel-Cyrille Ngonga Ngomo[1]

[1] Universität Leipzig, IFI/AKSW, PO 100920, D-04009 Leipzig
[2] Insight Center for Data Analytics, National University of Ireland, Galway
[3] Department of Computer Science, Universidad de Chile

Abstract. With this poster, we will present the Linked SPARQL Query (LSQ) dataset, which describes SPARQL queries taken from the logs of public endpoints. We introduce the initial four query logs that we have taken and the extraction process applied: the types of meta-data captured, how the data are modelled, what vocabularies we use, etc. The LSQ dataset currently contains 73 million triples describing 5.7 million query executions and is publicly available as Linked Data and through a SPARQL endpoint. We believe that by providing insights on how SPARQL is used in practice, the LSQ dataset could benefit areas of SPARQL research, including caching, benchmarking, usability, optimisations, etc.

1 Introduction

Public SPARQL endpoints collectively expose billions of facts and receive millions of queries per month. However, the maturity of SPARQL technology is still questionable: many endpoints have been found to suffer from service availability problems, or to exhibit non-standard behaviour such as silently returning partial results [2]. Evaluating SPARQL queries is computationally expensive for servers; in fact, it is known to be intractable [5]. Hence, *general* guarantees of efficiency cannot be made and hence, it is crucial to understand how SPARQL 1.1 is being used in practice, and to try to focus on those research questions with the highest potential for impact on real users, e.g., to look at what types of features, joins, etc., are most commonly used; what combinations lead to the slowest runtimes; why that is; and whether or not optimisations are possible for these common problematic cases. Research topics such as usability, caching, benchmarking, etc., could also benefit from having more information about how SPARQL is being used in practice and what sorts of workloads current SPARQL endpoints face.

To understand trends in how SPARQL is used in practice, perhaps the best place to look is the logs of various public SPARQL endpoints. The first such initiative along these lines was the USEWOD collection, which made a variety of such logs available [3]. However, these datasets are only accessible after having signed legal agreements, meaning that researchers and other interested parties are limited in their reuse of the data. Also, the format of the logs is ad-hoc, depending on their source.

In this poster,[4] we present the Linked SPARQL Query Log Dataset (LSQ). LSQ is a public Linked Dataset of SPARQL queries extracted from endpoint logs. The current

[4] This poster accompanies the accepted dataset paper [6].

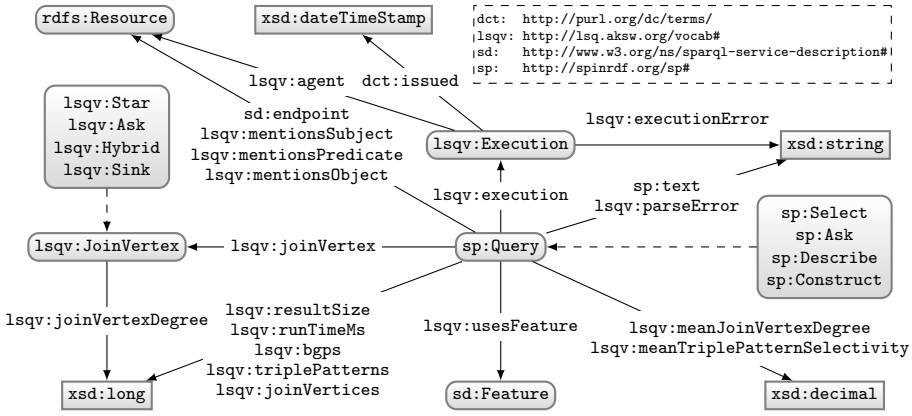

Fig. 1: LSQ data model (dashed lines indicate sub-class)

version consists of 73.2 million triples collected from four query logs for which we have obtained permission from the respective maintainers to make their content public. The LSQ dataset is available from `http://aksw.github.io/LSQ/`.

2 Data Model

Figure 1 summarises the LSQ data model, which reuses existing vocabularies where possible, combined with a custom LSQ vocabulary for new terms. The central class is `sp:Query`, where instances are typed as one of the subclasses: `sp:Select`, `sp:Ask`, `sp:Describe`, `sp:Construct`. We consider a query as unique to a log and potentially having multiple instances of `lsqv:Execution` (for each time the query was run) with a timestamp and a link to the agent who issued the request (identified using an encrypted IP). Query instances are linked to the query text (using `sp:text`) and to the originating endpoint (using `lsqv:endpoint`). To help make the dataset as general as possible, we also attach a complete SPIN representation of the query to each query instance [4].

Given that the SPIN representation may involve deep nesting, to make querying LSQ more convenient and efficient, we provide shortcut triples to indicate the SPARQL query features used in each query. These triples link query instances (with the predicate `lsqv:usesFeature`) to instances of `sd:Feature`. We enumerate a comprehensive list of such feature instances in our vocabulary, including `lsqv:Filter`, `lsqv:Optional`, `lsqv:SubQuery`, etc. We also provide shortcuts to the IRIs mentioned in a query with `lsqv:mentionsSubject`, `lsqv:mentionsPredicate` and `lsqv:mentionsObject`, making it easy to find queries mentioning a given resource.

In addition to the query structure, we also provide generic *structural statistics* [1] about the static query including the number of Basic Graph Patterns (`lsqv:bgps`) and the number of triple patterns (`lsqv:triplePatterns`). We also provide *data-driven statistics* [1] (incl. the number of results returned and the query runtime) about the execution of the query. Since such data are not typically provided by the logs, we generate these statistics by running the query locally against an offline copy of the

corresponding version of the dataset in question. Of course, the resulting statistics may differ to those that occurred during the original execution logged by the public endpoint and are rather intended as a guide (i.e., they are simply provided "as is").

With respect to Linked Data compatibility, we ensure that all query instances and executions are identified with dereferenceable URIs. Our data model also re-uses class and property terms from established external vocabularies, including SPIN, DC Terms and SPARQL Service Descriptions. Finally, with respect to external links, LSQ provides links to every URI mentioned in a query. We also provide a public SPARQL endpoint.

3 Current Query Logs

The current version of LSQ consists of queries extracted from four query logs as follows (see Table 1 for the corresponding statistics):

DBpedia is a broad encyclopaediac knowledge-base extracted from Wikipedia. A SPARQL endpoint is available at `http://dbpedia.org/sparql` powered by a Virtuoso instance. The DBpedia query log we have currently obtained spans from April 30, 2010 to July 20, 2010 (these queries refer to DBpedia v.3.5.1). The log records over 1.7 million query executions.

Linked Geo Data (LGD) contains a collection of spatial Linked Datasets that have been extracted from Open Street Map. The data are accessible via a public SPARQL endpoint at `http://linkedgeodata.org/sparql`, which uses Virtuoso as a back-end. The Linked Geo Data (LGD) query log spans from November 24, 2010 to July 6, 2011. The log records over 1.6 million query executions.

Semantic Web Dog Food (SWDF) is a community effort to generate a Linked Dataset about papers, presentations and people participating in top Semantic Web related conferences and workshops. The dataset is accessible through a SPARQL endpoint at `http://data.semanticweb.org/sparql` through a Sesame interface. The Semantic Web Dog Food (SWDF) log spans from May 16, 2014 to November 12, 2014 and records over 1.4 million query executions.

British Museum (BM) provides a Linked Data representation of an online collection containing records of more than 3 million artefacts. A SPARQL endpoint is accessible at `http://collection.britishmuseum.org/sparql` with an OWLIM/-GraphDB back-end. The log we have acquired spans from November 8, 2014 to December 1, 2014 and contains over 800 thousand query executions.

For potential consumers of LSQ, it is important to note that a high percentage of the millions of query executions recorded came from a small number of high-volume agents. Our goal with LSQ is to make details of the queries and their executions available "as is". Dealing with the issue of "agent skew" depends on what LSQ is to be used for, and thus is at the discretion of the LSQ consumer, who may wish to use data provided about agents and timestamps to, e.g., pre-filter the data. We refer to [6] for more details.

4 Poster Discussion

The current version of LSQ comprises 73 millions triples describing 5.7 million query executions from the four query logs introduced earlier. In the medium term, our plan

Table 1: Statistics of the datasets over which queries from the logs were executed

DATASET	TRIPLES	SUBJECTS	PREDICATES	OBJECTS	CLASSES
DBpedia	232,536,510	18,425,128	39,672	65,184,191	244
LGD	1,032,026,569	238,509,864	30,882	492,282,120	1,113
SWDF	294,870	30,856	185	93,051	126
BM	1,359,400	483,877	27	684,733	1

is to grow this dataset to include further logs, and thus we would like to use the poster session to discuss with endpoint maintainers the possibility of including their logs in the dataset. We would also be interested to discuss with SPARQL engine vendors the possibility of offering the option to produce logs directly in the LSQ format.

We would also like to speak with researchers working on SPARQL about the possible ways in which the LSQ dataset could benefit their work. We believe that LSQ could have applications to design benchmarks that better reflect the use of SPARQL in practice, to identify which SPARQL features (or combinations thereof) are most in need of optimisation, to empirically validate works on caching (by looking at sequences of query executions based on timestamps, or at the skew in terms of demand for data about popular resources), to inspire new research on usability or query relaxation (looking at how users behave: what mistakes they make, how they refine queries, etc.). We would also like to discuss with the community their ideas for what sorts of research questions or applications they might have in mind for the LSQ dataset, which may influence its design and provision, or may lead us to design new interfaces over the dataset.

In summary, by offering feedback on how SPARQL is used in practice, we believe that the LSQ dataset has the potential to help guide the development of resilient and reliable SPARQL systems operating under realistic work loads.

ACKNOWLEDGEMENTS: This work was supported in part by the German Ministry for Finances and Energy under the SAKE project (Grant No. 01MD15006E), by Science Foundation Ireland (SFI) under Grant No. SFI/12/RC/2289, by the Millennium Nucleus Center for Semantic Web Research under Grant No. NC120004 and by Fondecyt Grant No. 11140900.

References

1. G. Aluç, O. Hartig, M. T. Ozsu, and K. Daudjee. Diversified stress testing of RDF data management systems. In *ISWC*, pages 197–212, 2014.
2. C. B. Aranda, A. Hogan, J. Umbrich, and P. Vandenbussche. SPARQL web-querying infrastructure: Ready for action? In *ISWC*, pages 277–293, 2013.
3. B. Berendt, L. Hollink, V. Hollink, M. Luczak-Rösch, K. Möller, and D. Vallet. Usage analysis and the web of data. *SIGIR Forum*, 45(1):63–69, 2011.
4. H. Knublauch, J. A. Hendler, and K. Idehen, editors. *SPIN – Overview and Motivation*. W3C Member Submission, 22 February 2011.
5. J. Pérez, M. Arenas, and C. Gutierrez. Semantics and complexity of SPARQL. *ACM TODS*, 34(3), 2009.
6. M. Saleem, I. Ali, A. Hogan, Q. Mehmood, and A.-C. Ngonga Ngomo. LSQ: The Linked SPARQL Queries dataset. In *ISWC*, 2015. (to appear).