Accepted Manuscript

LDQL: A query language for the Web of Linked Data

Olaf Hartig, Jorge Pérez

Please cite this article as: O. Hartig, J. Pérez, LDQL: A query language for the Web of Linked Data, *Web Semantics: Science, Services and Agents on the World Wide Web* (2016), http://dx.doi.org/10.1016/j.websem.2016.10.001

LDQL: A Query Language for the Web of Linked Data☆

Olaf Hartig[a,∗], Jorge Pérez[b]

[a]Department of Computer and Information Science (IDA), Linköpings Universitet, SE-581 83 Linköping, Sweden
[b]Department of Computer Science, Universidad de Chile, Beauchef 851, Santiago - 8370456, Chile

Abstract

The Web of Linked Data is composed of tons of RDF documents interlinked to each other forming a huge repository of distributed semantic data. Effectively querying this distributed data source is an important open problem in the Semantic Web area. In this paper, we propose LDQL, a declarative language to query Linked Data on the Web. One of the novelties of LDQL is that it expresses separately (i) patterns that describe the expected query result, and (ii) Web navigation paths that select the data sources to be used for computing the result. We present a formal syntax and semantics, prove equivalence rules, and study the expressiveness of the language. In particular, we show that LDQL is strictly more expressive than all the query formalisms that have been proposed previously for Linked Data on the Web. We also study some computability issues regarding LDQL. We first prove that when considering the Web of Linked Data as a fully accessible graph, the evaluation problem for LDQL can be solved in polynomial time. Nevertheless, when the limited data access capabilities of Web clients are considered, the scenario changes drastically; there are LDQL queries for which a complete execution is not possible in practice. We formally study this issue and provide a sufficient syntactic condition to avoid this problem; queries satisfying this condition are ensured to have a procedure to be effectively evaluated over the Web of Linked Data.

1. Introduction

In recent years an increasing amount of structured data has been published and interlinked on the World Wide Web (WWW) in adherence to the Linked Data principles [2]. These principles are based on standard Web technologies. In particular, (i) the Hypertext Transfer Protocol (HTTP) [3] is used to access data, (ii) HTTP-based Uniform Resource Identifiers (URIs) [4] are used as identifiers for entities described in the data, and (iii) the Resource Description Framework (RDF) [5] is used as data model. Then, any HTTP URI in an RDF triple presents a *data link* that enables software clients to retrieve more data by looking up the URI with an HTTP request. The adoption of these principles has lead to the creation of a globally distributed dataspace: the *Web of Linked Data*.

The emergence of the Web of Linked Data makes possible an *online execution* of declarative queries over up-to-date data from a virtually unbounded set of data sources, each of which is readily accessible without any need for implementing source-specific APIs or wrappers. This possibility has spawned research interest in approaches to query the Web of Linked Data as if it was a single (distributed) database. For an overview on techniques proposed to execute queries over Linked Data on the WWW refer to [6].

While there does not exist a standard language for expressing such queries, a few options have been proposed in the research literature. In particular, a first strand of research focuses on extending the scope of the RDF query language SPARQL [7] such that an evaluation of SPARQL queries over Linked Data on the WWW has a well-defined semantics [8–12]. A second strand of research focuses on navigational languages [13, 14]. Although these approaches have different motivations, a commonality of all these proposals is that the definition of query-relevant regions of the Web of Linked Data and the definition of query-relevant data within the specified regions are mixed; as a result, in their queries, users cannot specify one without affecting the other.

☆This paper is an extended and revised version of [1].
∗Corresponding author
  Email address: `olaf.hartig@liu.se` (Olaf Hartig)