Chapter 17
SPORTAL:
Profiling the Content of Public SPARQL Endpoints

Ali Hasnain
INSIGHT Centre for Data Analytics, National University of Ireland, Ireland

Qaiser Mehmood
INSIGHT Centre for Data Analytics, National University of Ireland, Ireland

Syeda Sana e Zainab
INSIGHT Centre for Data Analytics, National University of Ireland, Ireland

Aidan Hogan
Center for Semantic Web Research, University of Chile, Chile

ABSTRACT

Access to hundreds of knowledge bases has been made available on the Web through public SPARQL endpoints. Unfortunately, few endpoints publish descriptions of their content (e.g., using VoID). It is thus unclear how agents can learn about the content of a given SPARQL endpoint or, relatedly, find SPARQL endpoints with content relevant to their needs. In this paper, the authors investigate the feasibility of a system that gathers information about public SPARQL endpoints by querying them directly about their own content. With the advent of SPARQL 1.1 and features such as aggregates, it is now possible to specify queries whose results would form a detailed profile of the content of the endpoint, comparable with a large subset of VoID. In theory it would thus be feasible to build a rich centralised catalogue describing the content indexed by individual endpoints by issuing them SPARQL (1.1) queries; this catalogue could then be searched and queried by agents looking for endpoints with content they are interested in. In practice, however, the coverage of the catalogue is bounded by the limitations of public endpoints themselves: some may not support SPARQL 1.1, some may return partial responses, some may throw exceptions for expensive aggregate queries, etc. The authors' goal in this paper is thus twofold: (i) using VoID as a bar, to empirically investigate the extent to which public endpoints can describe their own content, and (ii) to build and analyse the capabilities of a best-effort online catalogue of current endpoints based on the (partial) results collected.

DOI: 10.4018/978-1-5225-5191-1.ch017

1. INTRODUCTION

Linked Data aims at making data available on the Web in an interoperable format so that agents can discover, access, combine and consume content from different sources with higher levels of automation than would otherwise be possible (Heath & Bizer, 2011) . The envisaged result is a "Web of Data": a Web of structured data with rich semantic links where agents can query in a unified manner -across sources- using standard languages and protocols. Over the past few years, hundreds of knowledge bases with billions of facts have been published according to the Semantic Web standards (using RDF as a data model and RDFS and OWL to provide explicit semantics) following the Linked Data principles.

As a convenience for consumer agents, Linked Data publishers often provide a SPARQL endpoint for querying their local content (Jentzsch, Cyganiak, & Bizer, 2011) . SPARQL is a declarative query language for RDF in which graph pattern matching, disjunctive unions, optional clauses, dataset construction, solution modifiers, etc., can be used to query RDF knowledge bases; the recent SPARQL 1.1 release adds features such as aggregates, property paths, sub-queries, federation, and so on (Harris, Seaborne, & Prud'hommeaux, 2013). Hundreds of public endpoints have been published in the past few years for knowledge bases of various sizes and topics (Buil-Aranda, Hogan, Umbrich, & Vandenbussche, 2013; Jentzsch et al., 2011). Using these endpoints, clients can receive direct answers to complex queries using a single request to the server.

However, it is still unclear how clients (be they human users or software agents) should find endpoints relevant for their needs in the first place (Buil-Aranda et al., 2013; Paulheim & Hertling, 2013). A client may have a variety of needs when looking for an endpoint, where they may, for example, seek endpoints with data:

1. About a given resource, e.g., *MICHAEL JACKSON*;
2. About instances of a particular type of class, e.g., *PROTEINS*;
3. About a certain type of relationship between resources, e.g., *DIRECTS-MOVIE*;
4. About certain types of values associated with resources, e.g., *RATING*;
5. About resources within a given context or with specific values, for example, *CRIMES WITH LOCATION U.K. IN YEAR 1967 or RAT GENES AND DISEASE STRAINS*;
6. A combination of one or more of the above.

Likewise, a client may vary in how they are best able to specify these needs: some clients may only have keywords; others may know the specific IRI(s) of the resource, class or property they are interested in; some may be able to specify concrete queries or sub-queries that they wish to answer.

We argue that a service offering agents the ability to find relevant public endpoints on the Web would serve as an important part of the SPARQL querying infrastructure, enabling ad-hoc discovery of datasets over the Web. However, realising such a service over the current SPARQL infrastructure on the Web is challenging. Looking at the literature (in particular, works on the related problem of federated querying (Acosta, Vidal, Lampo, Castillo, & Ruckhaus, 2011; Harth et al., 2010; Hasnain et al., 2014, 2016; Quilitz & Leser, 2008; Schwarte, Haase, Hose, Schenkel, & Schmidt, 2011), we can find two high-level approaches that have been investigated thus far: