SpEnD Portal: Linked Data Discovery using SPARQL Endpoints

Semih Yumusak*, Riza Emre Aras*, Elif Uysal*, Erdogan Dogdu†, Halife Kodaz‡ and Kasim Oztoprak*
*KTO Karatay University,Karatay, KONYA, TURKEY
Email: name.surname@karatay.edu.tr
†Cankaya University,ANKARA, TURKEY
Email: edogdu@cankaya.edu.tr
‡Selcuk University, Selcuklu, KONYA, TURKEY
Email: hkodaz@selcuk.edu.tr

***Abstract*—We present the project SpEnD, a complete SPARQL endpoint discovery and analysis portal. In a previous study, the SPARQL endpoint discovery and analysis steps of the SpEnD system were explained in detail. In the SpEnD portal, the SPARQL endpoints are extracted from the web by using web crawling techniques, monitored and analyzed by live querying the endpoints systematically. After many sustainability improvements in the SpEnD project, the SpEnD system is now online as a portal. SpEnD portal currently serves 1487 SPARQL endpoints, out of which 911 endpoints are uniquely found by SpEnD only when compared to the other existing SPARQL endpoint repositories. In this portal, the analytic results and the content information are shared for every SPARQL endpoint. The endpoints stored in the repository are monitored and updated continuously.**

Sparql Endpoints, Linked Data, Search Engines

I. Introduction

Semantic web [2] was designed to re-organize the web into a web of connected entities. By this way, the knowledge around the web becomes understandable not only by humans, additionally by computers. By using the semantic web standards, the linked data was designed to interlink information between different data sources. According to the latest LOD Cloud Diagram[1], there are 1163 datasets which are compliant with the criteria defined to be listed in the LOD Cloud. These linked data sources can be queried by using a specific language called SPARQL [4] (SPARQL Protocol and RDF [7] Query Language). The SPARQL language can be used to query remote data sources as soon as the dataset is served through an endpoint, namely SPARQL endpoint. Technically, a SPARQL endpoint is a web page that allows a remote user to send a SPARQL query and get the result in a predefined format. Like any other source in the web, these SPARQL endpoints are served in web servers which may come across with several connectivity and maintenance issues.

The accessibility of linked data is a major issue for the services consuming online linked data sources. Whereas DBpedia [6] provides the biggest cross-domain online linked data source, it has been the only comprehensive source for years. Online services consuming linked data sources are highly dependent on SPARQL endpoints as they are the most widely used access interfaces for linked data sources. These SPARQL endpoints are listed in 5 different repositories (LOD Cloud [9], LODStats, Datahub [3], Sparqles, SpEnD). According to [10], more than half of those data sources listed are not maintaned properly. Most of the SPARQL endpoints listed in those repositories may not be reached and goes permenantly offline. Here we present a live portal allowing linked data consumers to browse a complete list[2] of SPARQL endpoints relevant to their consumption needs. At the backend, all SPARQL endpoints are queried regularly to ensure that a healthy information is provided for the end users.

The SpEnD portal serves many new SPARQL endpoints together with the following information if available:

- Category information,
- SPARQL compliance,
- Statistical information (e.g. the number of triples contained),
- Other SPARQL endpoints in the same domain,
- Query performance indicators

The SpEnD crawling engine described in [10] is transformed into a background service, which provides a continuous indexing of available SPARQL endpoints. This service is further used to analyze and monitor discovered endpoints.

Starting from a primary crawling stage, SpEnD portal discovers, analyzes and serves online the SPARQL endpoints available all over the Web.

The complete software stack of the SpEnD project is available as open source in the project repository[3] and the online portal is accessible via SpEnD web site[4]. The services belonging to the SpEnD portal are explained below in two sections: (1) Backend service, (2) Web Interface and API.

II. Backend Service

The SpEnD endpoint repository is populated by a back-end service, which is explained in detail in [10]. The current back-end is designed to be fully autonomous on the discovery and analysis of SPARQL endpoints processes unlike the

http://www.spend.semihyumusak.com.tr

[1]http://lod-cloud.net

[2]The SPARQL endpoint repository includes discovered endpoints from the SpEnD project as well as the ones listed in the other repositories (Datahub.io, SPARQLES [8], LODStats [5], LOD Cloud).

[3]https://github.com/semihyumusak

[4]http://spend.semihyumusak.com.tr

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

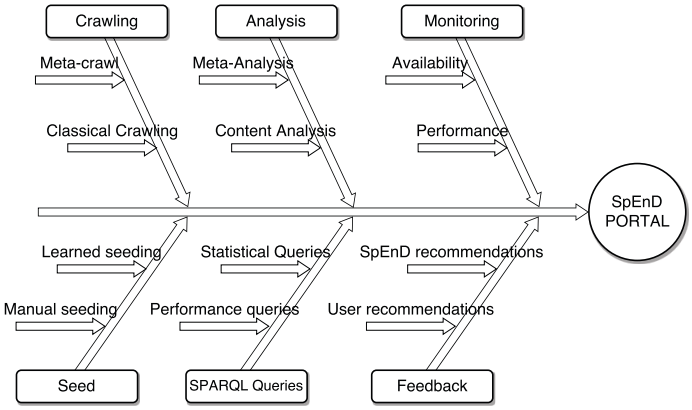

Fig. 1. SpEnD Workflow

earlier work. The current workflow of the complete system is summarized in Figure 1.

III. Web Interface and API

As seen in Figure 1, the SpEnD portal starts with a web crawling process. Then, the collected URLs are analyzed both on the content and meta-data. Finally, those collected and processed endpoints are monitored continuously.

The details of the processes are described in details as follows:

1) Seed: The seed words for the crawler are determined by collecting related web sites together with the domain knowledge. SPARQL endpoints are usually published as a standard web page depending on the data store behind (e.g. Virtuoso SPARQL endpoints, like Dbpedia[5] are always the same unless edited.)
2) Crawling: Crawling is the common method to index web pages all over the web. Other than classical crawling, this study additionally uses meta-crawling approach with the seeds described above.
3) SPARQL Queries: Benchmark statistical SPARQL queries [1] and performance queries [6] are imported into the system.
4) Analysis: A continuous analysis is performed by using the queries described above.
5) Feedback: Based on the analysis results and other user recommendations, the SPARQL endpoints list is updated.
6) Monitoring: A continuous monitoring of the SPARQL endpoint is performed. Finally, the list of SPARQL endpoints with the collected information is served through the SpEnD PORTAL.

The SpEnD back-end service feeds the SpEnD portal in terms of new SPARQL endpoints and statistical information about the currently listed SPARQL endpoints. In order to provide information for a linked data consuming service, a web site serves the data in proper formats. The SpEnD web site provides a search interface and also an API to let users query all discovered endpoints. All endpoints listed in this web site contains collected information to describe or statistically visualize the background information. The SPARQL endpoints and their details can be queried by using the search functionality of the web site, or by using the REST API service, which allows users to:

- Retrieve a list of all endpoints[7]
- Get detailed information about any endpoint[8]
- Search endpoints using keywords and subjects[9]

A. Dashboard

While collecting and analyzing the endpoints, many properties are collected (e.g. service type, subject, number of triples). Some of these are summarized in the dashboard screen as shown in Figure 2 and 3. In this dashboard screen, the latest status of the endpoints repository is summarized with a link to the relevant endpoints. The following properties about the SPARQL endpoints are categorized and visualized:

- Server type (e.g. Apache, nginx etc.)
- Content categories (based on LOD Cloud categories)
- Availability percentages
- Number of total triples per endpoints

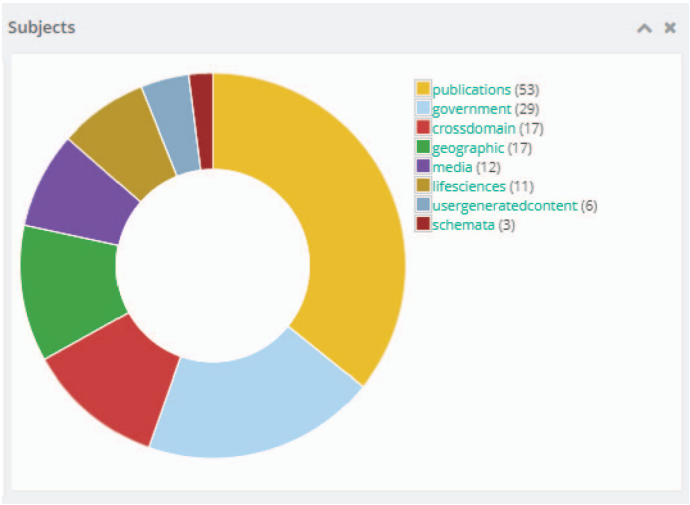

Fig. 2. Dashboard Screen-Subjects Chart

B. Endpoint Search and Details

The complete list of endpoints are available at `All Endpoints`[10] page. From the web portal, there are two ways of reaching to the endpoint details. The user may either visit the complete list and choose an endpoint, or search an endpoint with a word or phrase through `Search Endpoint`[11] page. The details page[12] contains the meta-analysis and content analysis results about an endpoint (Figure 4).

[5]http://dbpedia.org/sparql
[6]https://www.w3.org/2001/sw/DataAccess/tests/r2
[7]http://spend.semihyumusak.com.tr/api/endpoints?page={page number}
[8]http://spend.semihyumusak.com.tr/api/endpoint/{endpoint id}
[9]http://spend.semihyumusak.com.tr/api/search?=keyword{keyword}page={page}filter={su
[10]https://spend.semihyumusak.com.tr/reports/endpoints
[11]https://spend.semihyumusak.com.tr/search
[12]https://spend.semihyumusak.com.tr/endpoint/5971bfa9b891ecd560000486

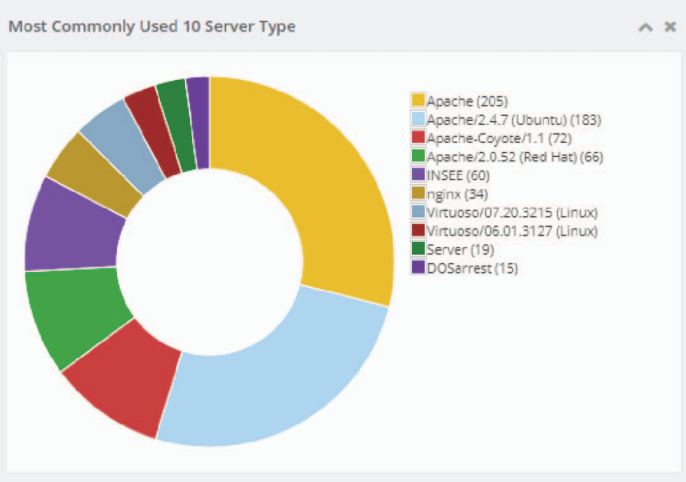

Fig. 3. Dashboard Screen-Server Types Chart

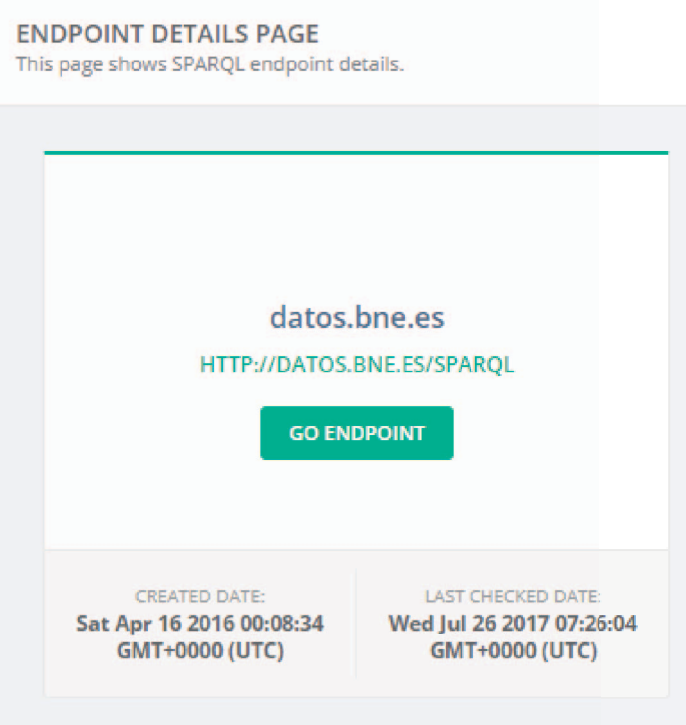

Fig. 4. Endpoint Details Screen

In this details page the following details of a SPARQL endpoint are listed, which is available in the Endpoint Details page shown in Figure 5:

- Subject (Category)
- Number of Triples
- Number of Entities
- Number of Distinct Resource URIs
- Number of Distinct Classes
- Number of Distinct Predicates
- Number of Distinct Subject Nodes
- Number of Distinct Object Nodes
- Server Type
- Source List

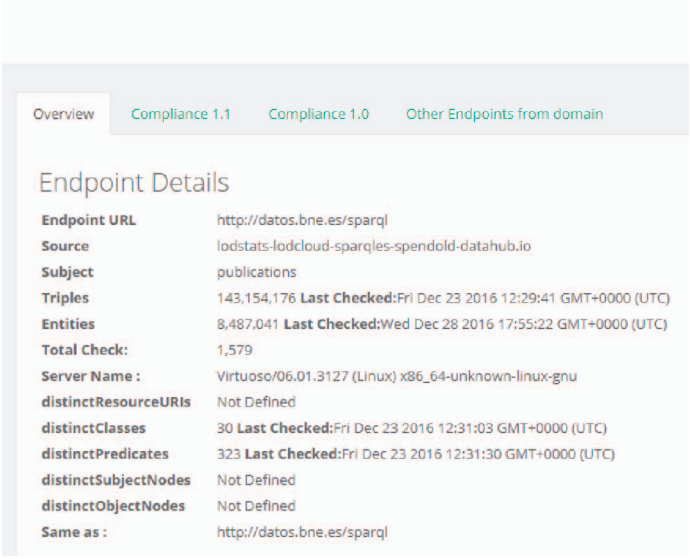

Fig. 5. Endpoint Details Screen

IV. Conclusion

In this study, the latest version of the SpEnD portal is presented, which has a complete workflow on SPARQL endpoints retrieval, analysis, and publishing. Endpoints listed in other projects are also available together with the ones discovered by the SpEnD back-end service. The current list of endpoints can be queried and used by linked data consumers.

References

[1] Alexander, K., Hausenblas, M.: Describing linked datasets-on the design and usage of void, thevocabulary of interlinked datasets. In: In Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09. Citeseer (2009)

[2] Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific american 284(5), 28–37 (2001)

[3] Bhardwaj, A., Bhattacherjee, S., Chavan, A., Elmore, A.J., Madden, S., Parameswaran, A.: Datahub: Collaborative data science & dataset version management at scale. In: In CIDR. Citeseer (2015)

[4] Buil-Aranda, C., Hogan, A.: SPARQL Web-Querying Infrastructure: Ready for Action? In: The Semantic WebISWC 2013. pp. 277–293. Springer Berlin Heidelberg (2013)

[5] Ermilov, I., Lehmann, J., Martin, M., Auer, S.: LODStats: The Data Web Census Dataset, pp. 38–46. Springer International Publishing, Cham (2016), http://dx.doi.org/10.1007/978-3-319-46547-0_5

[6] Lehmann, J., Isele, Robert and Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S.: DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web 6(2), 167–195 (2015)

[7] Miller, E.: An introduction to the resource description framework. Bulletin of the Association for Information Science and Technology 25(1), 15–19 (1998)

[8] Vandenbussche, P., Aranda, C., Hogan, A., Umbrich, J.: Monitoring the Status of SPARQL Endpoints. In: ISWC 2013 Demo. vol. 1380, pp. 3–6 (2013)

[9] Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and maintaining links on the web of data. The Semantic Web-ISWC 2009 pp. 650–665 (2009)

[10] Yumusak, S., Dogdu, E., Kodaz, H., Kamilaris, A., Vandenbussche, P.Y.: SpEnD: Linked Data SPARQL Endpoints Discovery Using Search Engines. IEICE Transactions on Information and Systems, E100-D(4), 758–767 (2017)