Dataset Dashboard – a SPARQL Endpoint Explorer

Petr Křemen, Lama Saeeda, Miroslav Blaško, and Michal Med

Czech Technical University in Prague, Praha, Department of Cybernetics, Knowledge-based Software Systems Group, Czech Republic
http://kbss.felk.cvut.cz
{petr.kremen,saeedla1,blaskmir,medmicha}@fel.cvut.cz

Abstract. We present the Dataset Dashboard, a SPARQL endpoint exploration tool that helps to understand the dataset structure, as well as relationships to other datasets. The tool is based on the notion of a dataset descriptor which describes some characteristic of a dataset. Currently, the tool offers descriptors for basic class/property statistics, spatial information, temporal information, as well as advanced dataset summarization. The tool currently registers over 200 SPARQL endpoints and named graphs inside the SPARQL endpoints.

Keywords: Linked Data · SPARQL · Dataset Descriptor

1 Introduction

When trying to understand unknown RDF datasets[1], linked data consumers needs to judge suitability of the dataset for his/her scenario in an efficient way. However, the dataset exploration itself is a non-trivial and time-consuming task for many reasons. First problem lies in the data – although many RDF datasets are cataloged, the metadata gathered about these datasets are typically not put in sync with the actual dataset content on the regular basis, or describe provenance metadata and do not reflect the dataset content at all. Second important problem lies in the accessibility of the data, which is often limited by (un)availability/latency of SPARQL endpoints, or efficiency of the underlying hardware.

In this paper we aim at the problem of dataset exploration by users with technical background who need to become familiar with unknown datasets (e.g. linked data designers who link to external sources, data journalists who explore a new linked open data source, etc.). As an example, let's consider a linked data designer who needs to find out, whether a new and unknown dataset is suitable for integration with another dataset (s)he is designing. For this purpose, a bunch of exploratory SPARQL queries need to be posed by the designer to become familiar with the *vocabulary used in the dataset*, *mutual interlinks between classes*, their significance, as well as the spatial (e.g. which spatial area it covers) or temporal (which temporal interval it spans) scope of the dataset.

The introduced Dataset Dashboard tool aims at making the process of understanding an unknown RDF dataset more efficient. For this purpose, the tool shows several

[1] In this paper, we only consider RDF datasets available as subsets of actual snapshots of SPARQL endpoints content.

interactive visualizations of general content-level characteristics of the dataset, including a filterable dataset summary, basic class and property statistics, as well as spatial and temporal range of the dataset.

2 Dataset Dashboard Overview

Please refer to the RDF(S) specifications [6],[5] for related notions. We consider an RDF graph of the form $g = \{\langle s_i, p_i, o_i\rangle\}$, where for each triple $\langle s_i, p_i, o_i\rangle$, s_i (resp. p_i, o_i) is its subject (resp. predicate, object). A *dataset descriptor* is an RDF graph $\delta(g)$ constructed from g which is *easier to interpret and visualize* than g itself. Function $\delta : 2^g \rightarrow 2^g$, is the actual *dataset descriptor function*. The main purpose of the dataset dashboard is to compute and visualize descriptors of various types.

Let's introduce basic features of the Dataset dashboard[2], which are later presented in Section 3 on an example.

Summary Schema Widget The basic summarization descriptor provided by the Summary Schema Widget is the SPO Summary, schematically described as

$$\delta_{SPO}(g) = \{\langle c_1, p, c_2\rangle | \{\langle x, \mathtt{rdf:type}, c_1\rangle, \langle y, \mathtt{rdf:type}, c_2\rangle, \langle x, p, y\rangle\} \subseteq g\}$$

Additionally, frequencies of the summary triples are computed and visualized by the weight (thickness) of the corresponding graph edge. The widget provides also an enhanced version of the SPO Summary descriptor which shows instance-level links to other datasets. The technique to compute enhanced descriptors is currently under review at the Journal of Web Semantics.

To explore large summaries, two filtering options are offered – **content-based filtering** allows to exclude concrete classes/properties from the summary graph, while **weight-based filtering** allows to exclude all edges with weight below a threshold, i.e. filtering out triple patterns which are not frequent enough in the dataset.

Spatial Widget Spatial part of the data is represented by geographical objects, using GeoSPARQL[3]. Spatial Widget extracts GeoSPARQL geometries and displays them in a map, schematically

$$\delta_{GeoSPARQL}(g) = \{\langle o, p, g\rangle | \langle o, p, g\rangle \in g, \text{for } p \in \{\mathtt{geosparql:asWKT}, \mathtt{geosparql:asGML}\}\}$$

The widget recognizes both Well-Known Text (WKT) geometry and Geography Markup Language (GML) geometry. All types of features found in the dataset are listed in the pop-up menu next to the map, where the user may choose which features to show in the map. Point, line and polygon geometries are supported. Moreover, points are represented as markers with pop-up bubble containing the link to the RDF resource representing the feature.

[2] Available on-line at `https://onto.fel.cvut.cz/dataset-dashboard`

[3] `http://www.opengeospatial.org/standards/geosparql`, cit. 1.6.2018

Temporal Widget Temporal widget shows the temporal coverage of the dataset, computed as minimum and maximum time points occurring in the dataset. The computation itself considers the structured temporal information as well as temporal information extracted from non-structured texts. For the unstructured temporal information, it retrieves the textual literals and performs a natural language processing analysis to extract the time information. To perform this step, the Stanford temporal tagger (SUTime)[4] which is part of the StanfordCoreNLP[5] tool is used. The pipeline for computing the temporal descriptor can be found in [10].

Dataset Dashboard Bookmarking In order to share a particular dataset dashboard across the community, bookmarkable dataset dashboard URIs are offered.

3 Example

Let's consider a dataset g_{exp} about parcels, buildings, floors and land use which is a part of the dataset maintained by the Prague Institute of Planning and Development[6]. Comparing to the original, g_{exp} is limited to the data for Prague Centre, which results in approx. 1.8 mil. triples.

Basic capabilities of the dataset dashboard are shown in Figure 1. The figure shows an overview of g_{exp} in terms of the SPO summary provided by the Summary Schema widget. The full SPO summary contains 13 RDFS classes and 25 RDF properties. The dataset dashboard allows to filter out some classes/properties from the schema. Additionally, edges can be filtered according to their weight. Thus, the figure excludes two properties from the view (`rdfs:label`, `ddo:has-published-dataset-snapshot`) and shows only edges with weight at least 12000 (adjustable by the slider control above the SPO summary graph).

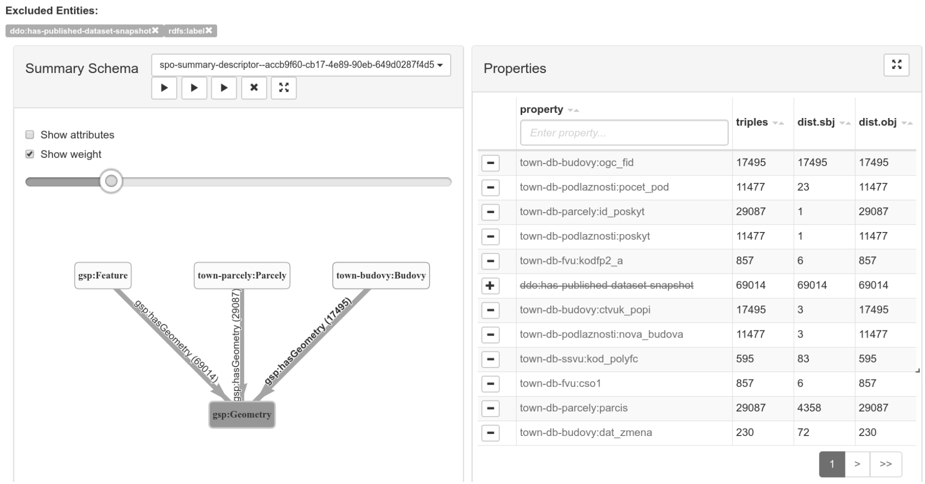
Spatial and temporal context of the dataset is shown in Figure 2. The GeoSPARQL descriptor shows geometries attached functional land use (instances of the RDFS class `town-fvu:VyuzitiPloch`). Dereferencable link of one of the geometries[7] is shown in a tooltip. The temporal descriptor shows the temporal range of the dataset. This dataset provides temporal information only inside the properties `town-parcely:dat_vznik` and `town-budovy:dat_vznik`, `town-budovy:dat_zmena`, which denote the creation/change dates of the data change record. So in this case, and contrary to the GeoSPARQL descriptor, the extracted temporal knowledge refers rather to the creation of the data than to the actual content. The dashboard of this example is shown at the following URL
`https://onto.fel.cvut.cz/dataset-dashboard/#/dashboard/online?endpointUrl=http://onto.fel.cvut.cz:7200/repositories/ipr_datasets&graphIri=http://onto.fel.cvut.cz/ipr-datasets/resource/town-plan`

[4] `https://nlp.stanford.edu/software/sutime.html`

[5] `https://stanfordnlp.github.io/CoreNLP/`

[6] `http://en.iprpraha.cz`, cit. 1.6.2018

[7] `http://onto.fel.cvut.cz/ontologies/town-plan/pvp_fvu_p/geometry/70/2018-01-29T14:36:24.178617`



Excluded Entities:

ddo:has-published-dataset-snapshot ✖ rdfs:label ✖

Summary Schema: spo-summary-descriptor--accb9f60-cb17-4e89-90eb-649d0287f4d5

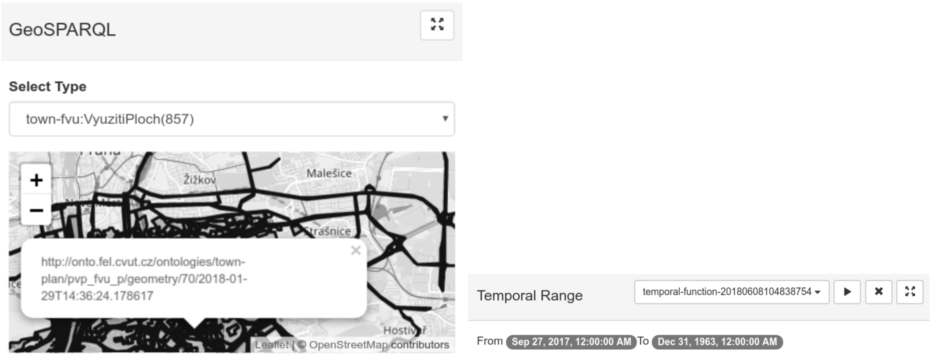
□ Show attributes

☑ Show weight

Properties

	property	triples	dist.sbj	dist.obj
−	town-db-budovy:ogc_fid	17495	17495	17495
−	town-db-podlaznosti:pocet_pod	11477	23	11477
−	town-db-parcely:id_poskyt	29087	1	29087
−	town-db-podlaznosti:poskyt	11477	1	11477
−	town-db-fvu:kodfp2_a	857	6	857
+	~~ddo:has-published-dataset-snapshot~~	69014	69014	69014
−	town-db-budovy:ctvuk_popi	17495	3	17495
−	town-db-podlaznosti:nova_budova	11477	3	11477
−	town-db-ssvu:kod_polyfc	595	83	595
−	town-db-fvu:cso1	857	6	857
−	town-db-parcely:parcis	29087	4358	29087
−	town-db-budovy:dat_zmena	230	72	230

Fig. 1: An example SPO summary and basic statistics widget.



GeoSPARQL

Select Type

town-fvu:VyuzitiPloch(857)

http://onto.fel.cvut.cz/ontologies/town-plan/pvp_fvu_p/geometry/70/2018-01-29T14:36:24.178617

Temporal Range: temporal-function-20180608104838754

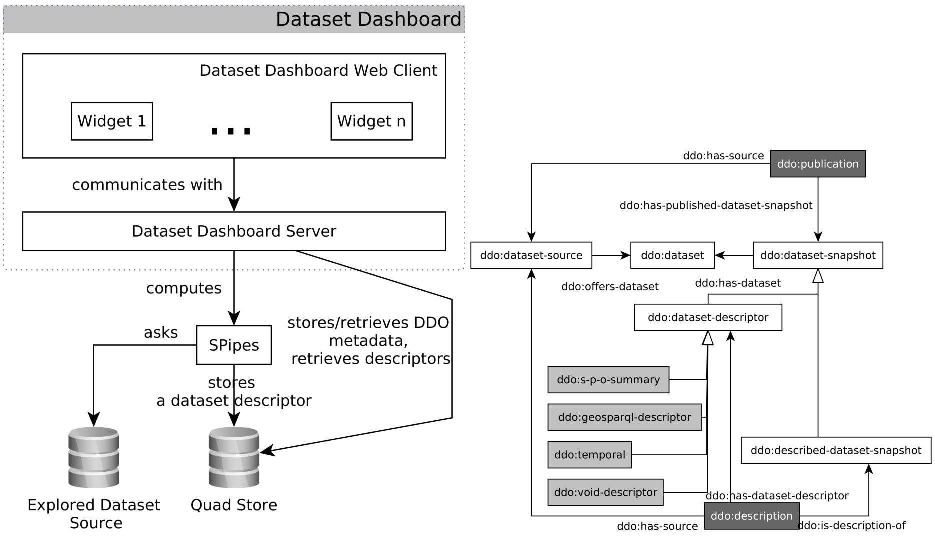
From Sep 27, 2017, 12:00:00 AM To Dec 31, 1963, 12:00:00 AM

(a) (b)

Fig. 2: Example temporal/spatial descriptors of the dataset.

4 Technical Background

Dataset dashboard is a client-server application[8], as shown in Fig. 3a. The client side serves mainly as a container for various *widgets*. Each widget is a separate React component, which is able to visualize a dataset descriptor of one or more types (see Section 4.1). The server takes care of proper creation and consumption of Dataset Descriptor Ontology metadata (see Section 4.1). In order to implement a new visualization, one needs to implement a visual widget component and just insert it in the `DatasetDashboardController`, along with the other widgets. Upon request, the widget gets from the Server the result of one of the predefined SPARQL queries. For more complicated scenarios, one needs to also implement an *SPipes* (see Section 4.2) pipeline that computes the particular descriptor.



(a) Dataset Dashboard Architecture

(b) Dataset Description Ontology Core

4.1 Dataset Descriptor Ontology

Dataset Descriptor Ontology (DDO) [4] is a domain ontology[9] for describing the process of dataset description, backed by the Unified Foundational Ontology (UFO) [8]. Simplified view on the ontology core is depicted in Figure 3b. The fundamental notion of the ontology is the notion of a ddo:dataset, being an endurant entity which can have different ddo:dataset-snapshots in time. Each dataset is available through some

[8] `https://kbss.felk.cvut.cz/gitblit/summary/?r=dataset-dashboard.git`, cit. 8.6.2018

[9] `http://onto.fel.cvut.cz/ontologies/ddo`, cit. 4.6.2018

ddo:dataset-source to which particular ddo:dataset-snapshots can be deployed by an event of ddo:dataset-publication. Upon an event of describing a dataset source (a ddo:description), a new ddo:dataset-descriptor is created as a description of a ddo:described-dataset-snapshot. An example of a ddo:dataset-snapshot is a particular content of a SPARQL endpoint.

4.2 SPipes

SPipes is a custom implementation of the SPARQLMotion engine[10]. It allows for creation RDF processing pipelines, consisting of *modules* (nodes) and their edges *dependencies* (edges). A dependency says that first the source module is executed and its resulting RDF graph together with global variable bindings is passed to the target module. Then the target module is executed. The language is very well integrated with SPARQL[13]. It allows reusing variables from SPARQL expressions in modules and vice versa. It can be accessed through REST interface that is generated from definitions of pipelines. Currently, SPipes are used to compute the SPO summaries and temporal descriptors.

5 Related Work

An overview of exploration tools can be found in [3] and a more recent one in [9]. Let's discuss some of the tools in detail.

LODeX [2] is a tool (no on-line demo available any more) offering visualization similar to the SPO summaries. *Comparing to Dataset Dashboard*, information about frequent triple patterns of the dataset is not graphically visualized (and thus cannot be used for filtering the summary).

LODSight [7] visualizes SPO summaries, in a similar way the Dataset Dashboard does. *Comparing to Dataset Dashboard*, LODSight does not allow to filter the SPO summary by the classes occurring in the data.

Linked Data Visualization Wizard [1] is an on-line tool[11] detecting presence of several types of information inside a SPARQL endpoint, including temporal information, spatial information, statistical data or SKOS. in addition to detecting presence of the information, the tool also shows examples of data for each type of information and the query used to generate them. *Comparing to Dataset Dashboard*, graphical visualization of the spatio-temporal characteristics and dataset summaries is missing.

Linked Geo Data Browser [11] allows to dynamically generate a faceted search based on linked data geographically located inside a region selected by user in a map. *Comparing to Dataset Dashboard*, it does not use the GeoSPARQL vocabulary and is bound to the LinkedGeoData dataset[12].

Facete [12] is a tool providing faceted search over a SPARQL endpoint, presenting the results in a map.Map4rdf[13] is a similar tool for GeoSPARQL-compliant data. *Comparing to Dataset Dashboard*, both tools seem to support only point geometries.

[10] `http://sparqlmotion.org/`, cit. 3.6.2018

[11] `http://semantics.eurecom.fr/datalift/rdfViz/apps`, cit. 11.7.2018

[12] `http://linkedgeodata.org`, cit. 11.7.2018

[13] `http://oegdev.dia.fi.upm.es/map4rdf/`, cit. 1.6.2018

Furthemore, comparing to the all mentioned tools, Dataset Dashboard provides a comprehensive view of the dataset under exploration by combining different descriptors into a single dashboard. Also, it stores the computed dataset descriptors[14] and provide persistent identifiers, for efficient sharing of the view over the dataset.

6 Evaluation

We conducted a preliminary survey about usefulness of the tool among three IT experts (a PhD student in the semantic web field, a linked data expert and a semantic web developer) not involved in the system design and development. As a part of the survey they had to explore three datasets (a SKOS vocabulary about labor law[15]), a complex dataset about EU television content[16]), a dataset with geospatial and temporal information about urban development in Prague[17]) not known to them before and judge the benefits of using tool for their use-case. Although the experts were not provided with any information about the tool beyond its SPARQL endpoint URL, all of them were successful in describing what the topics of all three datasets are (mainly using the Summary Schema Widget). Two of them see the main advantage of the tool in support for subsequent SPARQL query formulation to the particular SPARQL endpoint. The third one sees its advantage in providing visualization revealing the complexity of the dataset. Two of the experts find dataset dashboard persistent links useful for sharing information about the datasets. On the other hand, for the purpose of dataset exploration, the experts miss visualization of data samples and more advanced data statistics.

7 Conclusions

We presented the Dataset dashboard – a tool for dataset exploration using different *dataset descriptors*. The tool currently registers over 200 SPARQL endpoints and named graphs inside the SPARQL endpoints and is currently used in two national research projects. Initial feedback by IT experts is motivating, revealing that the tool is useful for dataset exploration, as well as providing suggestions for future work.

In future, we aim at providing history tracking for computed descriptors, as well as introducing new descriptor types (e.g. data samples, as suggested by the survey) for spatial and temporal widgets.

Acknowledgements This work was partially supported by grants No. GA 16-09713S Efficient Exploration of Linked Data Cloud of the Grant Agency of the Czech Republic and No.SGS16/229/OHK3/3T/13 Supporting ontological data quality in information systems of the Czech Technical University in Prague.

[14] Currently only for SPO summary and temporal descriptors.

[15] `http://vocabulary.wolterskluwer.de/PoolParty/sparql/arbeitsrecht`, cit. 10.7.2018

[16] `http://lod.euscreen.eu/sparql`, cit. 10.7.2018

[17] The dataset presented in Section 3

References

1. Atemezing, G.A., Troncy, R.: Towards a linked-data based visualization wizard. In: ISWC 2014, 5th International Workshop on Consuming Linked Data (COLD 2014), 20 October 2014, Riva del Garda, Italy. Riva Del Garda, ITALIE (10 2014), `http://www.eurecom.fr/publication/4380`
2. Benedetti, F., Bergamaschi, S., Po, L.: LODeX: A tool for visual querying linked open data. In: CEUR Workshop Proceedings. vol. 1486 (2015)
3. Bikakis, N., Sellis, T.K.: Exploration and visualization in the web of big linked data: A survey of the state of the art. In: Palpanas, T., Stefanidis, K. (eds.) Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference, EDBT/ICDT Workshops 2016, Bordeaux, France, March 15, 2016. CEUR Workshop Proceedings, vol. 1558. CEUR-WS.org (2016), `http://ceur-ws.org/Vol-1558/paper28.pdf`
4. Blasko, M., Kostov, B., Kremen, P.: Ontology-based Dataset Exploration – A Temporal Ontology Use-Case. In: Proc. of the Intelligent Exploration of Semantic Data (IESD'16). Kode (2016)
5. Brickley, D., Guha, R.V.: RDF Schema 1.1 (feb 2014), `http://www.w3.org/TR/rdf-schema/`
6. Cyganiak, R., Wood, D., Lanthaler, M.: RDF 1.1 Concepts and Abstract Syntax. W3c recommendation, W3C (2014), `http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/`
7. Dudáš, M., Svátek, V., Mynarz, J.: Dataset summary visualization with LODsight. In: Lecture Notes in Computer Science. vol. 9341, pp. 36–40 (2015). https://doi.org/10.1007/978-3-319-25639-9_7
8. Guizzardi, G.: Ontological foundations for structural conceptual models. Ph.D. thesis, University of Twente, The Netherlands. (mar 2005), `http://doc.utwente.nl/50826/1/thesis{_}Guizzardi.pdf`
9. Klímek, J., Škoda, P., Nečaský, M.: Survey of Tools for Linked Data Consumption. Semantic Web (2018)
10. Saeeda, L., Kremen, P.: Temporal knowledge extraction for dataset discovery. In: CEUR Workshop Proceedings. vol. 1927 (2017)
11. Stadler, C., Lehmann, J., Höffner, K., Auer, S.: Linkedgeodata: A core for a web of spatial open data. Semant. web **3**(4), 333–354 (Oct 2012), `http://dl.acm.org/citation.cfm?id=2590208.2590210`
12. Stadler, C., Martin, M., Auer, S.: Exploring the web of spatial data with facete. In: Proceedings of the 23rd International Conference on World Wide Web. pp. 175–178. WWW '14 Companion, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2567948.2577022, `http://doi.acm.org/10.1145/2567948.2577022`
13. The W3C SPARQL Working Group: SPARQL 1.1 Overview. W3c recommendation (2012), `https://www.w3.org/TR/sparql11-overview/`