Dataset Dashboard
A SPARQL Endpoint Explorer

Petr Křemen
petr.kremenh@fel.cvut.cz
Motivation

• DCAT metadata inside data catalogs are mostly agnostic to the actual content of the dataset

• How to become familiar with the content of a dataset and help designing a content-oriented metadata of a dataset

• Linked datasets instead of Linked Data (containing Linked data)
Motivation

- quickly become familiar with a SPARQL endpoint
  content from different general points of views

- RDF dataset summary (triple summary)
  - Enrichment with links to other datasets
  - Filterable by class/property facets

- Spatial information
  - GeoSPARQL

- Temporal information
  - Structured (dc:date, etc.)
  - Unstructured (literals)
**Dataset Descriptors**

**Dataset descriptor** of a dataset D is another dataset δ(D), which describes D and is easier to visualize.

- Basically any function of the *dataset content only*.

- *RDF summaries, geo extracts, temporal extracts*

```RDF
:John a :Person .
:mary a :Person .
:sue a :Person .
```

```RDF
[] rdf:subject Person ;
    rdf:predicate :loves ;
    rdf:object :Person ;
    dd:has-weight "2"^^xsd:int.
```
RDF Dataset Summary (Triple summary)

\[ \text{iff} \quad (\text{?sT} \rightarrow \text{sT}, \text{?p} \rightarrow \text{p}, \text{?oT} \rightarrow \text{oT}) \]

is a solution of

\[ [a \ ?sT] \ ?p \ [a \ ?oT] \]
For **untyped resources** find other datasets where they are typed using an **index of untyped resources**.

Faceted Filtering of Summaries

Excluded Entities:
ddo:has-published-dataset-snapshot, rdfs:label

Summary Schema
- Show attributes
- Show weight

Properties

<table>
<thead>
<tr>
<th>property</th>
<th>triples</th>
<th>dist.sbj</th>
<th>dist.obj</th>
</tr>
</thead>
<tbody>
<tr>
<td>ddo:has-published-dataset-snapshot</td>
<td>69014</td>
<td>69014</td>
<td>69014</td>
</tr>
<tr>
<td>town-db-budovy:ctvuk_popi</td>
<td>17495</td>
<td>3</td>
<td>17495</td>
</tr>
<tr>
<td>town-db-podlaznosti:octav_pod</td>
<td>11477</td>
<td>3</td>
<td>11477</td>
</tr>
<tr>
<td>town-db-podlaznosti:poskyt</td>
<td>11477</td>
<td>1</td>
<td>11477</td>
</tr>
<tr>
<td>town-db-podlaznosti:nova_budova</td>
<td>11477</td>
<td>3</td>
<td>11477</td>
</tr>
<tr>
<td>town-db-ssvu:kod_polyfc</td>
<td>595</td>
<td>83</td>
<td>595</td>
</tr>
<tr>
<td>town-db-fvu:csol</td>
<td>857</td>
<td>6</td>
<td>857</td>
</tr>
<tr>
<td>town-db-parcely:parcis</td>
<td>29087</td>
<td>4358</td>
<td>29087</td>
</tr>
<tr>
<td>town-db-budovy:dat_zmena</td>
<td>230</td>
<td>72</td>
<td>230</td>
</tr>
</tbody>
</table>
Spatial Information

- GeoSPARQL
  - SpatialObject
    - has geometry
    - Feature ➔ Geometry ➔ asWKT ➔ Literal

1. List of **frequent features types**
2. Visualization of **features of the selected type**
Temporal Information

Temporal Range

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

- Compute range of times found in the dataset
  - Structured data
    - White-list of properties analysed from LOV cloud
  - Unstructured texts inside literals
    - Extracted using SUTime library

Comparison with some other Tools

- **LODEX** (No public demo)
- **LODSight** (http://rknown.vserver.cz/lodsight)
  - Only property filtering (not classes)
  - No Geo/Temporal data
- **Linked Data Visualization Wizard** (http://semantics.eurecom.fr/datalift/rdfViz/apps)
  - Summaries?
  - Temporal data (only structured ones)
  - Geo data (WGS84, not GeoSPARQL)
- **LGD Browser and Editor** (http://browser.linkedgeodata.org/)
  - No summaries, no temporal data
  - More suitable for GeoSPARQL data
User study

• 3 IT experts
  • PhD student in semantic web
  • Linked data expert
  • Ontology application developer

• Task:
  • Describe topic of 3 unknown datasets
    • WK Arbeitsrecht (SKOS vocabulary about work law) http://bit.ly/dd-iswc-1

• All three IT experts were successful in describing the content of previously unknown dataset using RDF summarization widget

• Two IT experts claim that they can use the tool for subsequent SPARQL query formulation to the endpoint.

• All three experts miss example resource visualization
Future Work

- History tracking for computed descriptors
- New descriptors types (e.g. SchemEx, RDFSummary, Geo vocabulary)

THANK YOU

https://github.com/kbss-cvut/dataset-dashboard