Knowledge-based Software Systems
Faculty of Electrical Engineering
Czech Technical University in Prague,
Czech Republic

Dataset Dashboard

A SPARQL Endpoint Explorer

Petr Křemen
petr.kremen@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***

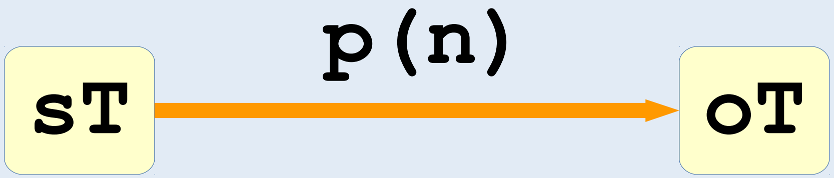
D

```
:John a :Person .
:mary a :Person .
:sue a :Person .
:John :loves :mary, :sue .
```

δ(D)

```
[] rdf:subject Person ;
   rdf:predicate :loves ;
   rdf:object :Person;
   dd:has-weight “2”^^xsd:int.
```

RDF Dataset Summary (Triple summary)

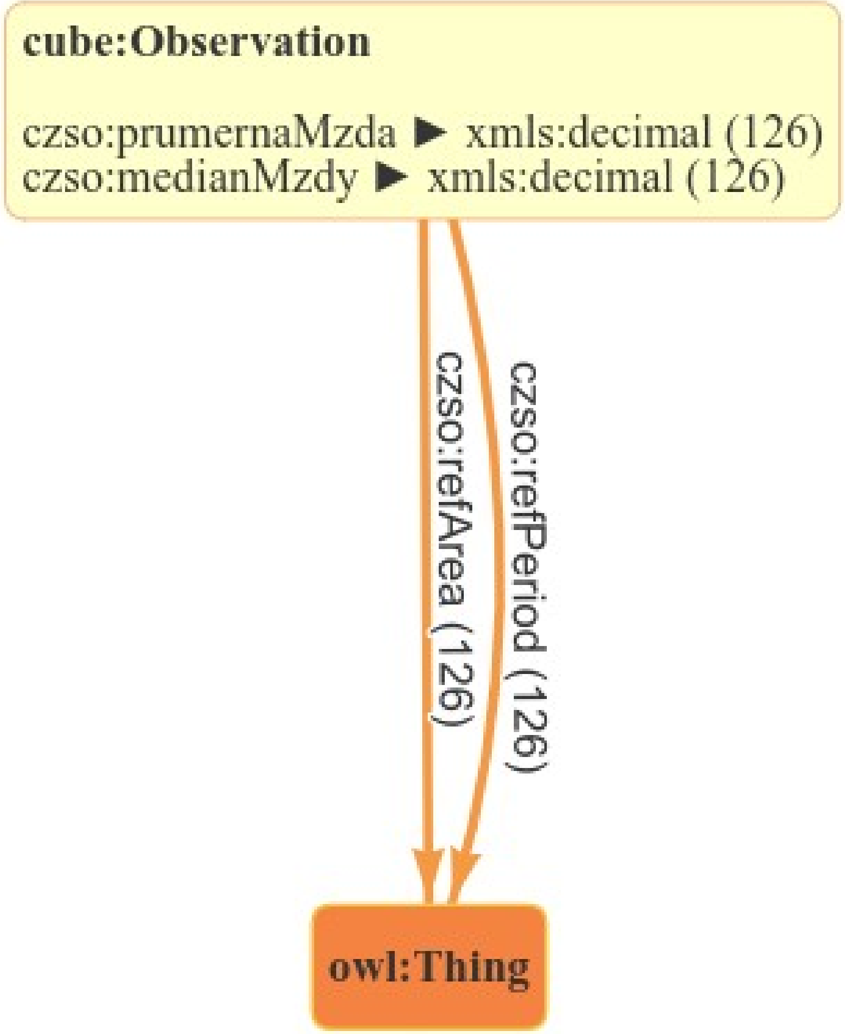


iff

(?sT→**sT**, ?p→**p**, ?oT→**oT**)

is a solution of

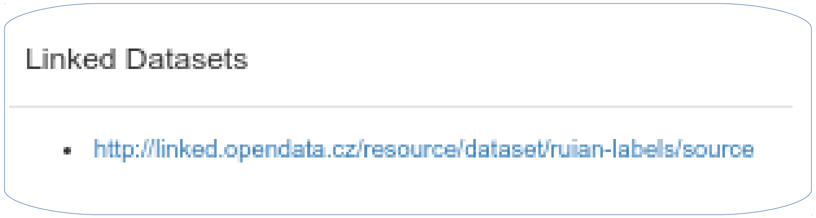
[**a** ?sT] ?p [**a** ?oT]



cube:Observation
czso:prumernaMzda ▶ xmls:decimal (126)
czso:medianMzdy ▶ xmls:decimal (126)
czso:refArea (126)
czso:refPeriod (126)
owl:Thing

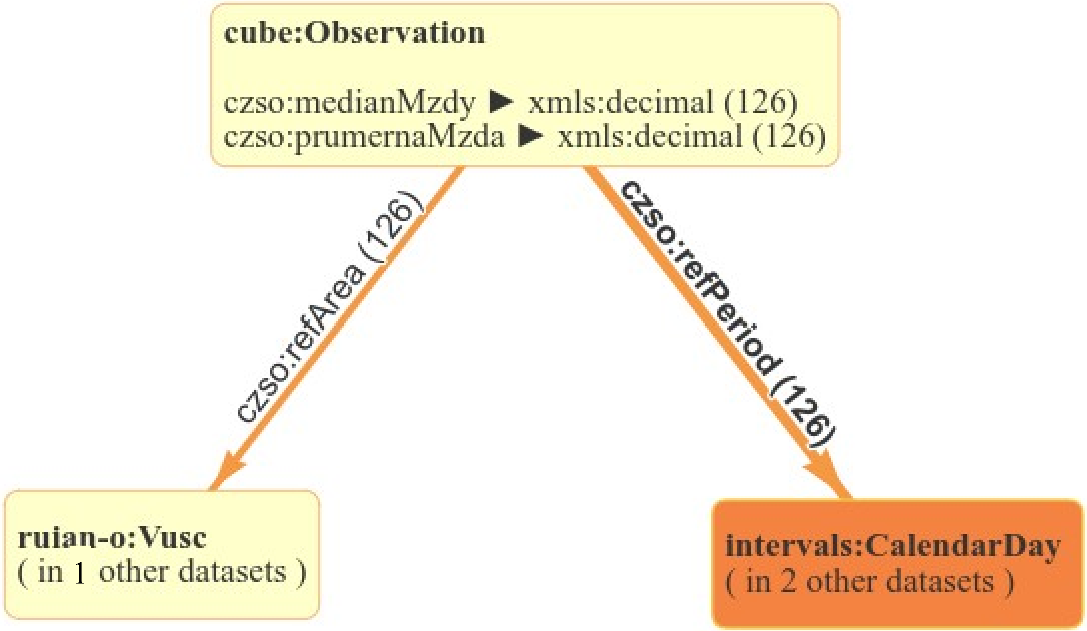
Richer RDF Dataset Summary

For **untyped resources** find other datasets where they are typed using an **index of untyped resources**.



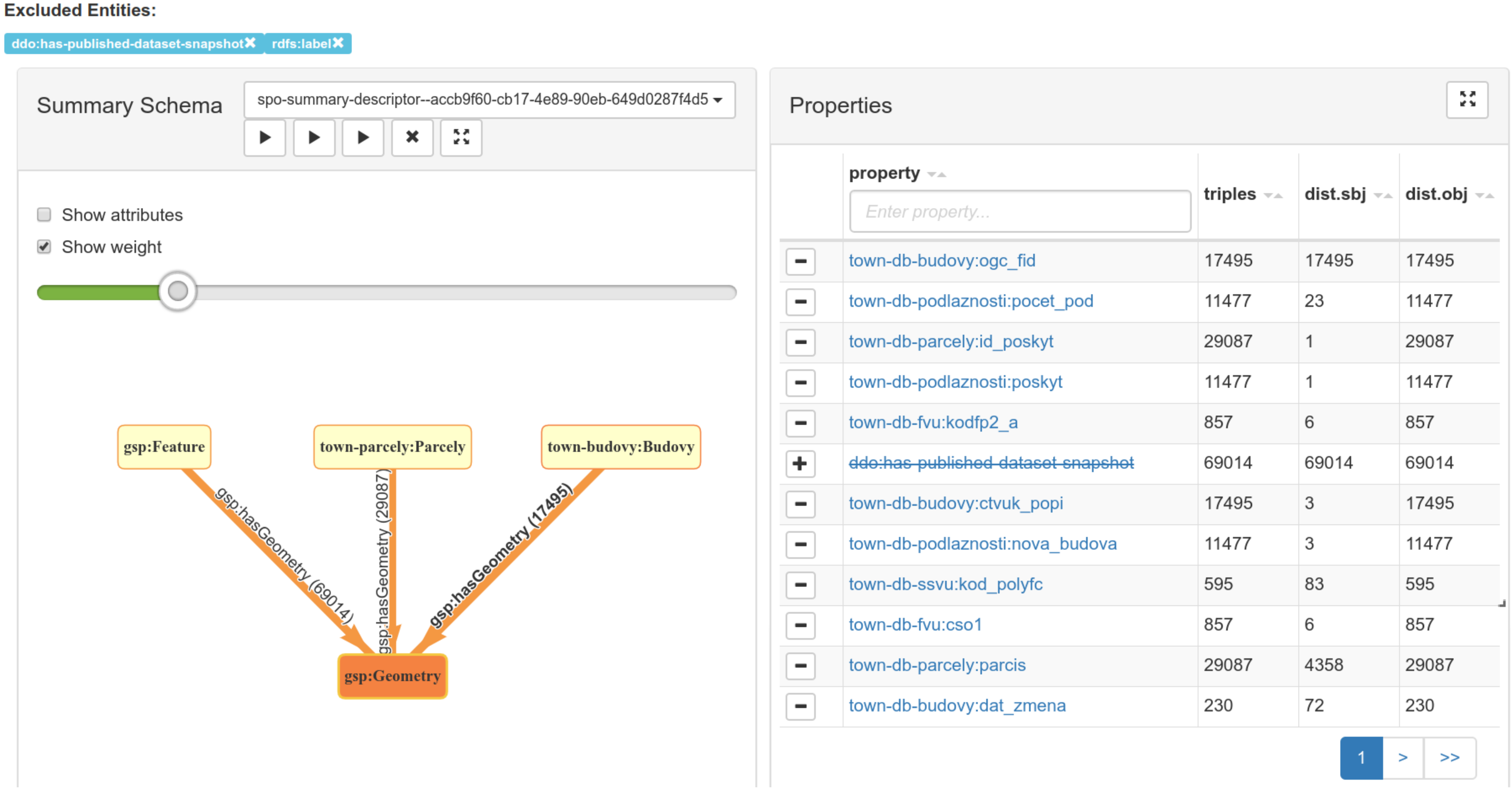
Linked Datasets

- http://linked.opendata.cz/resource/dataset/ruian-labels/source



P. Křemen, B. Kostov, M. Blaško, J. Klímek, and M. Nečaský. ***Towards Richer Dataset summaries***. Submitted to the Journal of Web Semantics in June 2018.

Faceted Filtering of Summaries



Excluded Entities:

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

Summary Schema

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

□ Show attributes

☑ Show weight

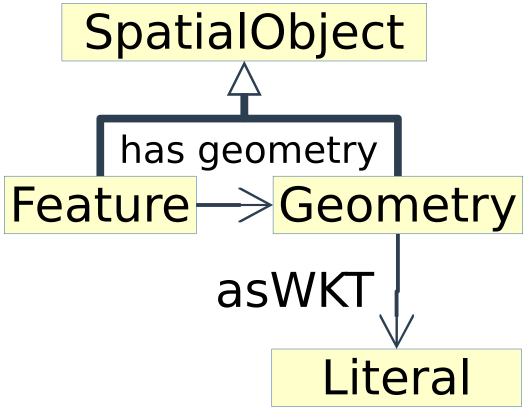
Properties

	property	triples	dist.sbj	dist.obj
	Enter property...			
−	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

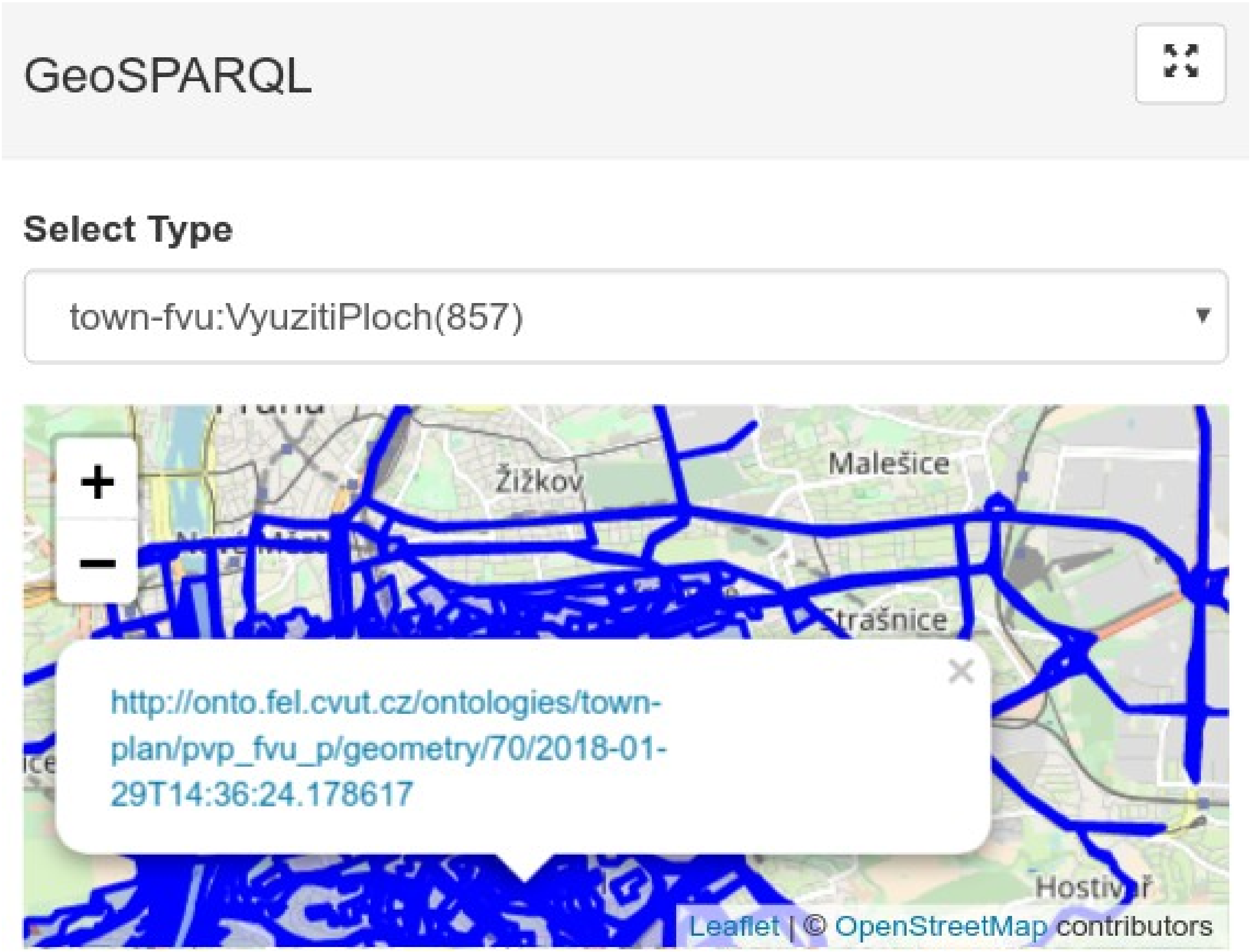
1 > >>

Spatial Information

- GeoSPARQL



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



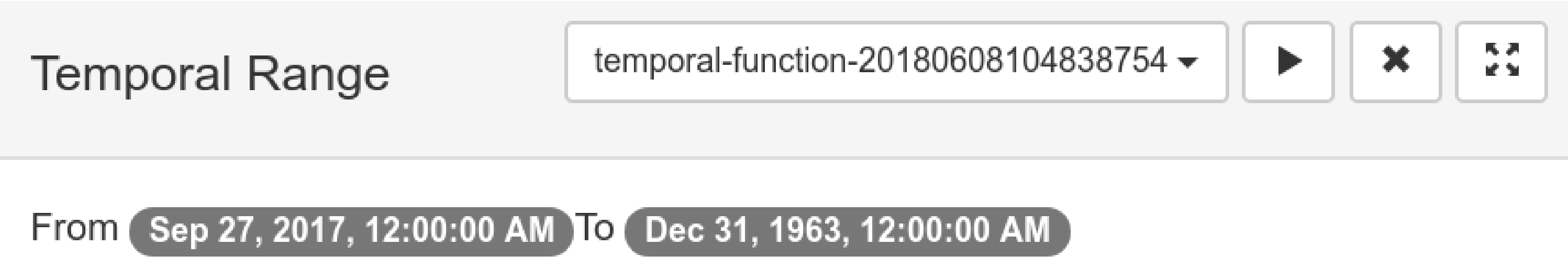
GeoSPARQL

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 Information



Temporal Range

temporal-function-20180608104838754 ▾

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

L. Saeeda, P. Křemen. ***Temporal knowledge extraction for dataset discovery***. In: CEUR Workshop Proceedings. vol. 1927 (2017)

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**
    - LOD Euscreen (EU TV content) **http://bit.ly/dd-iswc-2**
    - Urban planning dataset of Prague **http://bit.ly/dd-iswc-3**
- **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`