Development of St. Petersburg's linked open data site using Information Workbench

Dmitry Mouromtsev, Vitaly Vlasov, Mikhail Galkin, Vitaly Knyazev and Olga Parkhimovich
NRU ITMO, Russia
Outline

• Reasons why we do it
• Objectives and goals
• Methods and approaches
• Development process and tools
• Features
• Future work
Reasons

- Government has an opportunity to publish their data
- Society has a strong intention to control authorities’ work
Reasons

Current level of Open Government Data in Russia

Community
Reasons

- Government has an opportunity to publish their data
- Society has a strong intention to control authorities’ work
- One of the best fields to follow the Linked Open Data (LOD) concept
- Semantic technologies may help greatly
- Successful examples of such portals in the USA, the UK and other countries

- LOD + Semantic Web = ❤
Project Scope

Main goal:
• Provide a powerful and flexible platform to publish open government data

Primary objectives:
• Full support of RDF, triple stores and available SPARQL endpoint
• Big Data
• Different sources of datasets
• Visualizing the data in convenient forms
Approaches

1\textsuperscript{st} approach – processing the budget of Saint Petersburg

• \(\sim\)100 MB RDF data
• Google App Engine
• Trying out some algorithms to work with triples
• Manual visualizing
Approaches

14.11.2013

14th FRUCT Mikhail Galkin NRU ITMO
Approaches

2nd approach is based on The Information Workbench by fluid Operations. ([http://www.fluidops.com/](http://www.fluidops.com/))

What is the Information Workbench?

- A powerful tool that supports a large amount of SW formats (RDF, N3, Turtle, TriG, etc)
- Has built-in Sesame store, SPARQL endpoint and visualizing methods
- Satisfies the requirement about different sources of data by introduction of the Data Provider concept.

* source: [http://www.fluidops.com/information-workbench/](http://www.fluidops.com/information-workbench/)
Approaches

What is the Data Provider?

- A program
- A module of the IWB
- Connects to some source (local, online) and processes the data to the database according to its purpose
- Updates the data automatically and regularly
- Versatile processing methods - up to the developer
Approaches
Implementation

• Separate committees → large number of web-sites with data
• Roads repairs dataset – for a prototype

http://gov.spb.ru/gov/otrasl/tr_infr_kom/tekobjekt/tek_rem/

Initially, we have a table:

<table>
<thead>
<tr>
<th>Street Name</th>
<th>Roadway area (thousands m²)</th>
<th>Pedestrian zone area (thousands m²)</th>
<th>Total (thousands m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name1</td>
<td>X1</td>
<td>Y1</td>
<td>X1+Y1</td>
</tr>
<tr>
<td>Name2</td>
<td>X2</td>
<td>Y2</td>
<td>X2+Y2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Implementation

Processing to triples:
URI = URL

- Street Name → uri:Name
- Roadway area → uri:Roadway
- Pedestrian zone area → uri:Pedestrian
- Total area → uri:Total

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>uri:name1</td>
<td>rdf:type</td>
<td>uri:Name</td>
</tr>
<tr>
<td>uri:name1</td>
<td>rdfs:label</td>
<td>“Name1”</td>
</tr>
<tr>
<td>uri:name1</td>
<td>uri:Roadway</td>
<td>X1</td>
</tr>
<tr>
<td>uri:name1</td>
<td>uri:Pedestrian</td>
<td>Y1</td>
</tr>
<tr>
<td>uri:name1</td>
<td>uri:Total</td>
<td>X1+Y1</td>
</tr>
</tbody>
</table>
Features

• Transforming text data to Linked Data
• An ontology is constructed automatically
• In case of roads dataset – binding street names to a map
• Data is accessible via SPARQL endpoint
• Huge opportunities for improving
  - Number of injuries
  - The most dangerous roads
  - Reliability of newly repaired roads
Features
Future work

Extensive way
More Data Providers for government web-sites since everyone is doing front-end differently

Intensive way
A universal algorithm.
• Specify URL and a title of the data we need
• Algorithm crawls through a web page, finds the data and processes it to the datastore
Future work

• A universal algorithm
• More datasets

**MORE** datasets

• Work with authorities to obtain the data that is already suitable for machine processing
• Emphasis on practical usage of Linked Data
Thank you!

http://ailab.ifmo.ru/
mikhgalkin@gmail.com
olya.parkhimovich@gmail.com