Semantic Web Challenge – Billion Triple Track

“An autosuggest service based on LOD backlinks”

Ioannis Papadakis, Michalis Stefanidakis
Ionian University, Corfu, Greece
Interlinking Triplestores

- Augmented and meaningful interlinking between the triplestores of the lod cloud is vital to the success of the linked-data movement.
- The existence of dereferencable URIs alone is not enough to integrate datasets and make them accessible to humans and machines.
- The vision of a highly interlinked lod cloud depends on the number of references from within a local triplestore to entity URIs that are introduced in remote ones.
LOD Registry Services

- The necessary tools should be provided that would facilitate the automated discovery of ‘useful’, remote URIs
  - Such tools would not only be employed by lod dataset providers, but also exploited by end-user applications wishing to provide transparent access to the underlying lod cloud
- Lod registry services are proposed as a set of semantic web services that are based on the notion of the 'Registry', which is actually a registry of all the native entity URIs that are defined within a local triplestore
The Registry

- The Registry should be part of a scalable architecture for the LOD cloud.
- Triplestores would be able to address queries to the registries of other triplestores in order to discover and use possibly useful remote entity URIs.
- The Registry would be enhanced with other interlinking information besides local entity URIs.
  - For example, such information could be the number of backlinks of each local entity URI.
    - Backlinks are the references to one entity URI of a local triplestore that are found within triples in remote triplestores.
A Registry-powered architecture for the LOD cloud
Updating Registries with backlinking information in the real LOD world

1. Insert triple

2. Backlinking notification

3. Add SPARQL endpoint of A to the repositories that reference B:r

A: SPARQL endpoint URL of A

B: Backlink service URL of B
Asking triplestores information about foreign entity URIs – Registry-based answer
How can agents discover ‘useful’ entity URIs within remote triplestores?
A registry-powered autosuggest application – The semantic web challenge case study

- Preprocessing of the BTC 2011 dataset
  - Identification of the provenance triplestores that appear as sources of BTC
  - Clustering unique entity URIs, according to their provenance triplestores
  - Discovery of URI labels that enable human user access to entity URIs
  - Identification of the backlinks of the entity URIs that had a label
A registry-powered autosuggest application – The semantic web challenge case study: Triplestores

Table 1: Top 10 Triplestores, ranked by number of URI references

<table>
<thead>
<tr>
<th>Triplestore</th>
<th>owned URIs in BTC dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://hi5.com">http://hi5.com</a></td>
<td>49,696,108</td>
</tr>
<tr>
<td><a href="http://bio2rdf.org">http://bio2rdf.org</a></td>
<td>9,760,153</td>
</tr>
<tr>
<td><a href="http://freebase.com">http://freebase.com</a></td>
<td>3,907,919</td>
</tr>
<tr>
<td><a href="http://dbpedia.org">http://dbpedia.org</a></td>
<td>3,574,727</td>
</tr>
<tr>
<td><a href="http://gov.uk">http://gov.uk</a></td>
<td>2,012,636</td>
</tr>
<tr>
<td><a href="http://loc.gov">http://loc.gov</a></td>
<td>1,811,393</td>
</tr>
<tr>
<td><a href="http://fu-berlin.de">http://fu-berlin.de</a></td>
<td>1,321,031</td>
</tr>
<tr>
<td><a href="http://scinets.org">http://scinets.org</a></td>
<td>1,184,674</td>
</tr>
<tr>
<td><a href="http://livejournal.com">http://livejournal.com</a></td>
<td>1,096,995</td>
</tr>
<tr>
<td><a href="http://dbtune.org">http://dbtune.org</a></td>
<td>1,047,444</td>
</tr>
</tbody>
</table>
A registry-powered autosuggest application – The semantic web challenge case study: Label predicates

Table 2: Predicates and label counts per predicate

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Label Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://open.vocab.org/terms/sortLabel">http://open.vocab.org/terms/sortLabel</a></td>
<td>58,519</td>
</tr>
<tr>
<td><a href="http://purl.org/dc/elements/1.1/title">http://purl.org/dc/elements/1.1/title</a></td>
<td>2,062,880</td>
</tr>
<tr>
<td><a href="http://www.fao.org/aims/aos/languagecode.owl#hasEnglishName">http://www.fao.org/aims/aos/languagecode.owl#hasEnglishName</a></td>
<td>7,642</td>
</tr>
<tr>
<td><a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#label">http://www.w3.org/1999/02/22-rdf-syntax-ns#label</a></td>
<td>6</td>
</tr>
<tr>
<td><a href="http://www.w3.org/2000/01/rdf-schema#label">http://www.w3.org/2000/01/rdf-schema#label</a></td>
<td>8,540,192</td>
</tr>
<tr>
<td><a href="http://www.w3.org/2004/02/skos/core#altLabel">http://www.w3.org/2004/02/skos/core#altLabel</a></td>
<td>1,517</td>
</tr>
<tr>
<td><a href="http://www.w3.org/2004/02/skos/core#prefLabel">http://www.w3.org/2004/02/skos/core#prefLabel</a></td>
<td>259,895</td>
</tr>
<tr>
<td><a href="http://zeitkunst.org/bibtex/0.1/bibtex.owl#title">http://zeitkunst.org/bibtex/0.1/bibtex.owl#title</a></td>
<td>963</td>
</tr>
</tbody>
</table>
The backlinking module satisfies requests for backlinks of specific local entity URIs within a triplestore.

The autosuggest module provides human-searchable access to the entity URIs of a triplestore. Moreover, it adds to the registry a searchable text index, built from the labels of the entities.
A registry-powered autosuggest application – The semantic web challenge case study: End-user application

- The proposed application provides the opportunity to address queries consisting of some characters to the entire BTC 2011 dataset or to a certain triplestore.
- The users are able to choose whether they want the resulting URIs to be ranked according to the number of their backlinks or, alphabetically ranked.

autosuggest application
Conclusions

- The selected approach is based on a modular architecture that can be easily adopted from existing LOD providers and capable of being extended with yet-to-come applications focusing on the enhancing of interlinking between LOD triplestores.
- The vision of a highly interlinked LOD cloud needs much further work in order to become a reality.
Future work

• Come up with more registry-based semantic web services
• Registry-based semantic web services should solve trust issues (do we grant everybody write-access to our Registry?)
• How do you determine the context of an entity URI within a remote triplestore’s Registry?
• How do you solve disambiguation issues regarding an entity URI within a remote triplestore’s Registry?
Thank you,

- Further information about the proposed work is publicly available in our research group’s homepage: [http://swrg.ionio.gr/](http://swrg.ionio.gr/)
- Source code will be soon available