Towards a Dynamic Linked Data Observatory

Tobias Käfer¹, Jürgen Umbrich², Aidan Hogan², Axel Polleres³

WWW2012 Workshop: Linked Data on the Web (LDOW2012)

¹) KARLSRUHE INSTITUTE OF TECHNOLOGY, GERMANY ²) DERI, NUI GALWAY, IRELAND ³) SIEMENS AG ÖSTERREICH, VIENNA, AUSTRIA
What’s this all about?

The Web

- Dynamic
  - Pages get created
  - Pages get updated
  - Pages get deleted
- Dynamicity causes problems
  - Cache freshness etc.
  - Studied and analysed

Aren’t we facing similar problems?
What’s this all about? (Cont’d)

The Web of Data

- Dynamic, too
  - Data gets created, updated, deleted
  - Vocabularies change, predicates are renamed
- Dynamicity influences…
  - Synchronisation of indexes
  - Smart caching of Linked Data content
  - Hybrid search engine architectures
  - …

→ Creation of a corpus to study the dynamics of Linked Data: The Dynamic Linked Data Observatory
Building blocks of a Dynamic Linked Data Observatory

Idea of what to monitor + Way of capturing the dimension of time

Means to create snapshots: LDspider + Bricks (for the sake of the metaphor) = The Dynamic Linked Data Observatory
We need an idea of what to monitor, but:

HOW TO GET A REPRESENTATION OF LINKED DATA?
Requirements for a representation of Linked Data and two candidates

- **Coverage**
  - Size
  - Diverse data providers
  - Balanced representation of data providers

- **Representativeness**
  - Study something people consider as LOD

---

**Genesis: Register dataset, meet requirements**

**Genesis: A crawl**

---

Towards a Dynamic Linked Data Observatory // TOBIAS KÄFER, Jürgen Umbrich, Aidan Hogan, Axel Polleres // LDOW 2012 @ WWW 2012

http://swse.deri.org/dyldo
Pros and cons of both datasets

**LOD/CKAN**

**Pros**
- Domains pass “quality control”
- Community validated

**Cons**
- Covers fewer domains* (133)
- Misses vocabularies
- Misses decentralised datasets like GoodRelations

**BTC2011**

**Pros**
- Covers more domains* (791)
- Empirically validated
- Includes vocabularies
- Includes decentralised datasets

**Cons**
- Influence of high-volume domains → unbalanced
- Misses 47.4% of LOD/CKAN domains

* pay-level domains (PLDs) to be precise
WHAT WOULD WE MISS BY CHOOSING EITHER OF THEM?

LOD/CKAN vs. BTC2011
What sites* would we miss, which would we get? (Top 10 statements)

More fine-grained analysis in the paper!

*pay-level domains (PLDs) to be precise
Our conclusion: a compromise

- Combination of CKAN/LOD-Cloud and BTC2011
  - Our sample:
    - 220 example URIs from the LOD-Cloud’s bubbles
    - 220 highest-ranked (PageRank) URIs from BTC2011*
    - Crawl from there to get a reasonably big seedlist

OUR MONITORING SETUP
Our setup

Download seedlist + Crawl = Published data:
- Seedlist
- The data itself
- access.log
- Frontier of the crawl after each hop

Taking into account RDF/XML, Turtle, RDFa, N-Triples, Nquads
The Dimension of Time: Sketch of our adaptive revisiting scheme (only for seedlist URIs)

URI changed between two visits

weekly

bi-weekly

quater-weekly
Summary / Q&A

Summary:
- Motivated Dataset Dynamics
- Contrasted CKAN/LOD and BTC2011
- Described our setup

Status quo:
- Close to launch (never been so close)
  - Expected: May 1
- Web page up
  - [http://swse.deri.org/dyldo](http://swse.deri.org/dyldo)
- Google Group up
  - [http://groups.google.com/group/dyldo](http://groups.google.com/group/dyldo)

Outlook
- Expected run-time: 1 year
- Elaborate on publishing issues
- Interpret data

Q&A
- What would be your use-case?
  - Does it need changes to our setup?
- How do you like our working definition of Linked Data?

Thanks for your attention!
This presentation is CC BY-SA

- Picture on title slide based on a picture by A. Sparrow
  http://www.flickr.com/photos/49937157@N03/
  - CC BY 2.0
- Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch.
  http://lod-cloud.net/
  - CC BY-SA
- Treasure hunting map by kruxmux
  http://www.flickr.com/photos/76476049@N00/3946522483/in/photostream
  - CC BY-NC 2.0
- Clock picture by millynet
  http://www.flickr.com/photos/millynet/134071210/lightbox/
  - CC BY-NC-SA 2.0
- Lens picture by Ben Cooper
  http://www.flickr.com/photos/cycleologist/1454436980/
  - CC BY-NC-SA 2.0
- Picture on last slide by http://www.flickr.com/photos/stevendepolo/
  - CC BY 2.0
Domination of large exporters in BTC:
One provider shapes overall characteristics

BTC2011 dataset

RDF from http://www.hi5.com in the BTC2011 dataset
Reasons for largest 10 PLDs in CKAN/LOD not appearing in BTC 2011

<table>
<thead>
<tr>
<th>PLD</th>
<th>ROBOTS</th>
<th>HTTP-401</th>
<th>HTTP-502</th>
<th>MIME</th>
<th>UNREACHABLE</th>
<th>OTHER</th>
</tr>
</thead>
<tbody>
<tr>
<td>linkedgeodata.org</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>concordia.ca</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rdfabout.com</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>unime.it</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>uriburner.com</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sudoc.fr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>viaf.org</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>europeana.eu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>moreways.net</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>uberblic.org</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Table 2: Reasons for largest ten PLDs in CKAN/LOD not appearing in BTC 2011.
Excursus: The PLD (pay-level domain)

- Pay money to a Top-level domain registrar
  → get a PLD

- Examples:
  - http://urq.deri.ie/
  - http://www.bbc.co.uk/programmes/b006ml0g

- Same notion, different name:
  - “Site” (Bray, WWW5, 1996)
  - “Top Private Domain” (Google Guava Libraries)