Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources

J. Umbrich, M. Hausenblas, A. Hogan, A. Polleres, S. Decker
Motivation (Situation)

In late 2009: we have over 100 open datasets providing over 13.7 billion RDF triples, interlinked by some 142 million links ...

... we can expect/have many changes (new resources, added links, updates)

courtesy of Richard Cyganiak and Anja Jentzsch
Motivation (Dataset Dynamics)

- Outcomes/Findings/Lessons learned from
  - LDC09: Dataset Change Manager/Watch-Dog demo
  - Dataset Dynamics Group

How does a dataset consumer learn when something has changed and what has changed?
Motivation (Dynamics in Web of Documents)

- Web of Documents
  - Change frequency of Web documents can be modeled as a Poisson Process [Cho at al. VLDB 2000]
  - Improvement of Estimators for uncertainty [Cho at al. ACM Journal 2001]
  - What is new on the Web? (search engine perspective) [Ntoulas et al., WWW2004]

- Applications
  - Web crawling and caching
  - Maintaining link integrity
  - replication and synchronisation
  Servicing of continuous queries
How dynamic is the Web of Data?
Detection Level: A Matter of Granularity

- **Document-centric**
  - HTTP Get body content

- **Entity-centric**
  - Entity-per-document change
  - (Global) Entity-change
    - Reuse of entities among sources (square)
    - wrt. to specific documents
      - e.g. per namespace, pay-level-domain (e.g. deri.ie)
## Change Detection Mechanism

- **Content monitoring**  
  fetching the entire content

- **HTTP caching**  
  (response header as of RFC 2616)

- **Notification**  
  active notification

<table>
<thead>
<tr>
<th></th>
<th>Content</th>
<th>HTTP</th>
<th>Notification</th>
</tr>
</thead>
<tbody>
<tr>
<td>availability</td>
<td>+</td>
<td>+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>reliability</td>
<td>+</td>
<td>+/-</td>
<td>unknown</td>
</tr>
<tr>
<td>costs</td>
<td>high</td>
<td>low</td>
<td>unknown</td>
</tr>
<tr>
<td>scalability</td>
<td>high</td>
<td>high</td>
<td>unknown</td>
</tr>
<tr>
<td>documents</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>entities</td>
<td>no</td>
<td>partiallyyes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Preliminary Results & Lessons learned
Experimental Setup

- **Dataset – Web data - (passive monitoring)**
  - weekly snapshots over 24 weeks (start Nov. 08)
  - 4 hop neighborhood from Tim Berners-Lee FOAF file
  - 550K RDF/XML docs, 3.3M unique entities

- **Change detection**
  - skolemise blank nodes within a document
  - pairwise comparison of statements by scanning sorted list

- **Detection level**
  - Document-centric
  - Entity-change (**pay-level-domain**)
Average Change Frequency

>60% is static

Entities more dynamic than documents

Average change frequency

<table>
<thead>
<tr>
<th>Category</th>
<th>Documents</th>
<th>Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>62.12%</td>
<td>68.88%</td>
</tr>
<tr>
<td>&lt;1 week</td>
<td>9%</td>
<td>52%</td>
</tr>
<tr>
<td>&gt;1 week ≤ 1 month</td>
<td>9%</td>
<td>24%</td>
</tr>
<tr>
<td>&gt;1 month ≤ 3 month</td>
<td>23%</td>
<td>23%</td>
</tr>
<tr>
<td>&gt;3 month ≤ 6 month</td>
<td>14%</td>
<td>59%</td>
</tr>
</tbody>
</table>
## Entity Changes per Document

### Legend:
- **U** = Update
- **A** = Add
- **D** = Delete

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>A</th>
<th>D</th>
<th>(UA</th>
<th>UD</th>
<th>AD)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U</strong></td>
<td>76.88 %</td>
<td>9.46 %</td>
<td>7.08 %</td>
<td>3.87 %</td>
<td>97.29 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A</strong></td>
<td>9.46 %</td>
<td>0.19 %</td>
<td>2.29 %</td>
<td>3.87 %</td>
<td>15.81 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>7.08 %</td>
<td>2.29 %</td>
<td>0.23 %</td>
<td>3.87 %</td>
<td>13.5 %</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fraction of Changes per Document

Only 10% of the content changes in 90% of the documents
Conclusion

- Web of Data has similar characteristics compared to the Web of Document concerning change frequencies
  - Change frequency of Web documents can be modeled as a Poisson Process [Cho at al. VLDB 2000]
- Document-centric change detection not helpful
  - different focus (entity-centric)
- Still needs more experimental verification
Future Work

- Large scale experiment over a long period
- Study of what and how much changes

- Dataset Dynamics Group:
  http://groups.google.com/group/dataset-dynamics

- Meet-up during WWW2010 on 29 April:
  http://esw.w3.org/Camps:LODCampW3CTrack#breakout

Questions?