Semantic Sitemaps

R. Cyganiak, H. Stenzhorn, R. Delbru, S. Decker, G. Tummarello
DERI Galway
A new Web (of Data)

- Old: documents for Web browsers
- New: structured data for mashups and application integration
- Key technology: RDF
Observation: Costs are shifting

- Access to RDF data was hard (DMOZ, MusicBrainz)
- Today: SPARQL protocol, Linked Data, Tabulator, …
- Today: More data (FOAF, Linking Open Data)
- Problem is no longer access but discovery
Towards a map of the new Web

• Swoogle
• SWSE
• Falcon-S
• Watson
• Sindice
3 challenges
1. Different access methods

Linked Data
RDF dumps
SPARQL endpoints
- GET to http://dbpedia.org/resource/Tenerife
- Dumps from http://downloads.dbpedia.org/
- SPARQL to http://dbpedia.org/sparql
- Same data everywhere
2. Crawl performance

- Toy servers, aggressive crawlers
- 1 request per second = 2.6M per month
- Geonames has 6M+ entities
- If a dump is available, how would a crawler know?
3. Provenance

- Is built-in feature of the Web (DNS)
- URI ownership, authoritative information
- Delegation of URI space not visible
Proposed solution
Semantic Sitemaps

- Publishers tell us where they have RDF data
- Based on Google’s Sitemap protocol
- Put a simple XML file on your server
Google’s Sitemap protocol

http://example.com/sitemap.xml

```
<urlset>
  <url>
    <loc>http://www.example.com/</loc>
    <lastmod>2008-01-01</lastmod>
    <changefreq>monthly</changefreq>
  </url>
  ... more ... 
</urlset>
```
Semantic Sitemaps

<urlset>
  ...
  <sc:dataset>

</sc:dataset>
</urlset>
Semantic Sitemaps

<urlset>
  ...
  <sc:dataset>
    <sc:linkedDataPrefix>
      http://dbpedia.org/resource/
    </sc:linkedDataPrefix>
  </sc:dataset>
</urlset>
Semantic Sitemaps

<urlset>
  ...
  <sc:dataset>
    <sc:linkedDataPrefix>
      http://dbpedia.org/resource/
    </sc:linkedDataPrefix>
    <sc:dataDumpLocation>
      http://downloads.dbpedia.org/dump.nt.gz
    </sc:dataDumpLocation>
  </sc:dataset>
</urlset>
Semantic Sitemaps

<urlset>
  ...
  <sc:dataset>
    <sc:linkedDataPrefix>
      http://dbpedia.org/resource/
    </sc:linkedDataPrefix>
    <sc:dataDumpLocation>
      http://downloads.dbpedia.org/dump.nt.gz
    </sc:dataDumpLocation>
    <sc:sparqlEndpointLocation>
      http://dbpedia.org/sparql
    </sc:sparqlEndpointLocation>
  </sc:dataset>
</urlset>
Semantic Sitemaps

?urlset>
 ... 
 <sc:dataset>
   <sc:linkedDataPrefix>
     http://dbpedia.org/resource/
   </sc:linkedDataPrefix>
   <sc:dataDumpLocation>
     http://downloads.dbpedia.org/dump.nt.gz
   </sc:dataDumpLocation>
   <sc:sparqlEndpointLocation>
     http://dbpedia.org/sparql
   </sc:sparqlEndpointLocation>
   <changefreq>monthly</changefreq>
 </sc:dataset>
</urlset>
More elements

- sc:datasetLabel: Name for the dataset
- sc:datasetURI: Hook for additional metadata
- sc:authority: Hook for identifying the publisher
- sc:sampleURI: Some representative URIs from the DS
- ...
Why XML?

- Conservative webmasters
- Simple
Sitemap discovery

http://domain/robots.txt

User-agent: *
Disallow:
Sitemap: sitemap.xml

http://domain/sitemap.xml

<urlset>
  ...
</urlset>
I. Different access methods

- Clients can choose between
  - sc:linkedDataPrefix
  - sc:dataDumpLocation
  - sc:sparqlEndpointLocation
2. Crawl performance

- Crawlers can discover and use RDF dump
- Experiment: Downloading and slicing Uniprot takes ~25h and can be parallelized
- Crawling Uniprot would take ~5 months
- Bottleneck moves from retrieval to indexing
3. Provenance

• Delegating and joining URI spaces with sc:subSitemap and sc:parentSitemap
• Describing the publisher with sc:authority
• URI space can be authoritatively served from a dump or SPARQL endpoint
Community and adoption

- Most large LOD datasets have a sitemap
- Supported by Sindice and SWSE
- Publishers are receptive
- They want a validator
- public-lod@w3.org mailing list
Next steps

- Updated draft
- Sitemap creator + validator
- Work on content descriptions (VOID)
Semantic Sitemaps …

• … are a proposal for better RDF discovery
• … allow publishers to announce their data
• … allow consumers to efficiently find it
• … have hooks for describing content and authority
richard@cyganiak.de