State of Play of OGC Web Services across the Web


Advanced Information Systems Laboratory (IAAA)
Universidad de Zaragoza, Spain
Outline

- Idea
- OWS Focused Crawler
- Results of April-May 2010
- Conclusions
Idea

- Searching OWS services in catalogues
  - Incomplete solution: voluntary registry
  - Does not guarantee validity of information

- Automated discovery of public OWS services using crawling techniques
  - Requires a focused OWS crawler

- Sources
  - Search engines
  - Geoportals
  - OGC catalogues
OWS Focused Crawler

- Design
  - Geoportals
  - Search engines
  - OGC Catalogues

- Challenges
  - XML Links
  - Lack of textual descriptions
  - OWS Exception reports
  - Links from Web applications
Results of April-May 2010

Questions that can be answered upon results?

- What is the size of public OWS in Europe?
- Do search engines cover the public OWS?
- Which is the most common specification?
- Which are the patterns of deployment?
- Where are the services found located?
The size of public OWS in Europe?

- **Services found**
  - 6,544

- **Estimated scale**
  - (6,684 – 5,757) CI 95%

- **Methodology**
  - Capture-recapture with 4 sources
Do search engines cover the public OWS?

- Search engines do not cover all the public OWS.
- Do we want to keep our services hidden?

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Found OWS</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo!</td>
<td>3,071</td>
<td>46.9%</td>
</tr>
<tr>
<td>Google</td>
<td>2,639</td>
<td>40.4%</td>
</tr>
<tr>
<td>Bing</td>
<td>1,761</td>
<td>26.9%</td>
</tr>
</tbody>
</table>

3,626 OWS found in Europe through search engines.

6,544 OWS found in Europe.
Which is the most common specification?

- Focus on portrayal services
- Low penetration of new standards
- Bad administration practices?
  - Many services running without operating on data
Which are the patterns of deployment?

- **Deployment data summary**

<table>
<thead>
<tr>
<th>Services per Host</th>
<th>Types per Service</th>
<th>Types per Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1\textsuperscript{st} quartile</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Median</td>
<td>2.00</td>
<td>2.00</td>
</tr>
<tr>
<td>Mean</td>
<td>11.55</td>
<td>7.30</td>
</tr>
<tr>
<td>3\textsuperscript{rd} quartile</td>
<td>6.00</td>
<td>5.00</td>
</tr>
<tr>
<td>Maximum</td>
<td>1,125.00</td>
<td>948.00</td>
</tr>
</tbody>
</table>

- **Simple services**
  - 50\% of hosts have 1 or 2 WMS
  - 50\% of servers serve only 1 or 2 map layers

- **Coexist with**
  - Service farms
  - Oversized services
Where are the active found services located?

- Cartogram: services vs. country size

- More services in:
  - Small/Medium sized countries north-central Europe
  - Large countries with decentralization (DE, ES, IT)

**ES bias:**
- Several service farms
- Search engines rank first results near from where the query is made
Conclusions

- Crawling offers an overview of the state of public OWS
  - It is possible to create a search engine from these results
  - But, it has technical challenges

- Crawling offers stakeholders “real-time” snapshots of the status of INSPIRE Network services

- Crawling offers valuable conclusions about current status of services, for example:
  - Focus on portrayal
  - Low penetration of recent OGC standards
  - Bad administration practices
  - Prevalence of simple services
Acknowledgement

This work has been partially supported by Spanish Government (projects “España Virtual” ref. CENIT 2008-1030, TIN2009-10971 and PET2008_0026), the Aragón Government (project PI075/08), the National Geographic Institute (IGN) of Spain, and GeoSpatiumLab S.L.