GeoSeer, A Global Search Engine for OGC Services

Jonathan Moules

INSPIRE Workshop, Ispra, July 2019
• GeoSeer
• Statistics
• What Works
• Problems
• Summary/Conclusions
GeoSeer
Search over **1.4 million** distinct spatial **WMS, WFS, WCS**, and **WMTS** layers hosted on over **190,000 services** from around the world.

**Search Tips**
- **British Columbia** or **geology Australia** or **rivers Asia** - Search spatially.
- **tree preservation 27700** - Include an EPSG code to prefer results in that projection.
- **arbol Spain** - Not all layers are in English, try the official language.
- **wcs Germany chart** - Search for a specific service (wcs, wmts, wfs, wms). We'd suggest only including one service in a search.
- **York -“New York“** - Standard search features are supported too. **Exact phrase searching** (wrap with "") and **exclude** (prefix with -)
- **wms (Alps | Himalayas)** - Boolean logic - by default GeoSeer puts an **AND** between all your search terms meaning **all terms are mandatory** in the result. You can use **OR** (with the pipe: | ) instead to expand the search results, and brackets () for grouping.

**Latest blog post:** **2019-04-09** - GeoSeer API Goes Live

Blog [RSS Feed](#)
# Portals as Seeds

<table>
<thead>
<tr>
<th>Portal Type</th>
<th>Number</th>
<th>Appx. Number of Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKAN</td>
<td>194 Instances</td>
<td>10,000*</td>
</tr>
<tr>
<td>CSW</td>
<td>151 Instances (481 – June 2019)</td>
<td>21,000</td>
</tr>
<tr>
<td><a href="http://opendata.arcgis.com">http://opendata.arcgis.com</a></td>
<td>-</td>
<td>800</td>
</tr>
<tr>
<td><a href="http://inspire-geoportal.ec.europa.eu">http://inspire-geoportal.ec.europa.eu</a></td>
<td>-</td>
<td>13,000+</td>
</tr>
<tr>
<td>GeoSeer</td>
<td>-</td>
<td>186,170</td>
</tr>
</tbody>
</table>

Table shows the **approximate** number of live distinct Services that can be found within a certain data-portal or type of data-portal (April 2019)

* Over a third of these are the US NOAA
+ Over half of these are one German Agency (www.geoportal.rlp.de)
GetCapabilities Documents

<?xml version="1.0" encoding="UTF-8"?>
  <Service>
    <Name>WPS</Name>
    <Title>Scottish Government - Marine Scotland Maps - Web Map Service</Title>
    <Abstract>An OGC compliant Web Map Server used to deliver spatial data to the National Marine Plan Interactive (NMPPI) web mapping application and associated services. If using this service in desktop GIS, you must ensure you have appropriate licences for any layers.</Abstract>
    <KeywordList>
      <Keyword>WFS</Keyword>
      <Keyword>WPS</Keyword>
      <Keyword>GEOSERVER</Keyword>
      <Keyword>MARINE</Keyword>
      <Keyword>SCOTLAND</Keyword>
      <Keyword>NPPC</Keyword>
      <Keyword>NATIONAL</Keyword>
      <Keyword>PLAN</Keyword>
    </KeywordList>
  </Service>
</GetCapabilities>
Postprocessing

- Data cleaning
- De-duplication
- Location Extraction
- Meaningless layer removal
- “Merging” layers (i.e. WMS + WFS)
What works
Data Portals

• GetCapabilities - Everything in “one” place
• They’re (kind of) curated
• API’s
  – Discovery Standards
  – Proprietary but reverse engineered
Discovery Standards are Good

- CSW Standard
- CKAN API

- Large deployments of these
“GetCapabilities” Document

- Title
- Name
- Abstract
- Metadata
- Keywords
- Bbox
- Etc.
Problems
Outdated lists

4 of 969 worked! (0.4%)

(and 2 of the 4 are the same!)
HTTP (insecure) vs HTTPS (secure)
“Registration Required”

- Data Portals
- OGC Services
Other Portals

• DKAN
  – No Discovery API

• Bespoke Portals
De-jure over De-facto

- ArcGIS “Geoportal”
- inspire-geoportal.ec.europa.eu – CSW
- (Not so much an issue for OGC Services)
Duplicates / Default Data

- Same data across many services
- One organisation had over 2 million layers!
- Common Default data
  - Spearfish
  - TOPP
  - SF
  - TIGER
  - etc
Metadata

• No-one bothers!
• Internal metadata is better
  – Less to crawl
  – No need to “glue” things together
  – High Reliability
  – Less bitrot
  – All in “one place”; i.e. adding to QGIS
Service Metadata

- `<Service>`
  - `<Name>`WMS`</Name>`
  - `<Title>`WMS`</Title>`
  - `<Abstract>`WMS`</Abstract>`
- `<KeywordList>`
  - `<Keyword>`WMS`</Keyword>`
- `<KeywordList>`
  - `<Keyword>`meteorology`</Keyword>`
  - `<Keyword>`atmosphere`</Keyword>`
  - `<Keyword>`climate`</Keyword>`
  - `<Keyword>`ocean`</Keyword>`
  - `<Keyword>`earth science`</Keyword>`

Metadata at the service level

Excludes default abstracts
Layer Metadata

metadata_url

<Name>mass_concentration_of_chlorophyll_a_in_sea_water</Name>
>Title>mass_concentration_of_chlorophyll_a_in_sea_water</Title>
<Abstract>mass_concentration_of_chlorophyll_a_in_sea_water</Abstract>

Service Types

WMS
WFS
WCS
WMTS

Layers %

No Keywords
No Abstract
No Keyword and No Abstract
No metadata_url

Metadata at the layer level
Layer Metadata - Abstracts

Table showing Layer percentages with meaningful abstracts, grouped by Service Type
“Meaningless” layers

- Title
  - No Title
  - Title is a random string of gibberish
- No Abstract or Keywords
- 237,382 “meaningless” layers removed for June 2019
Poorly Configured Servers

• Declared GetCapabilities_url is wrong
  − Missing scheme
  − Localhost,
  − Otherwise invalid
    731 services (~0.3%)

• Bad XML / Exceptions

• No Fee/Access Constraint info provided (~50%)

• INSPIRE – Of 28,078 services that mention “inspire”, 6,403 (~22%) don’t have an ExtendedCapabilities
Other things

- Unreliable Services
  - Offline during the crawl = not in the index
- Language barrier
Statistics
Number of Hosts, Services, and Layers per Country (URL domain).
Highlights the different national strategies in deployment.

.us, .gov, .edu merged
.scot, .wales, .uk merged
.brussels, .br merged
.int, .eu, .com, .net, .org Not included
Histograms

<table>
<thead>
<tr>
<th>Country</th>
<th># Hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>.us</td>
<td>905</td>
</tr>
<tr>
<td>.de</td>
<td>337</td>
</tr>
<tr>
<td>.ca</td>
<td>174</td>
</tr>
<tr>
<td>.fr</td>
<td>152</td>
</tr>
<tr>
<td>.uk</td>
<td>150</td>
</tr>
<tr>
<td>.nl</td>
<td>123</td>
</tr>
<tr>
<td>.es</td>
<td>107</td>
</tr>
<tr>
<td>.fi</td>
<td>94</td>
</tr>
<tr>
<td>.se</td>
<td>84</td>
</tr>
<tr>
<td>.no</td>
<td>74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country</th>
<th># Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>.us</td>
<td>49800</td>
</tr>
<tr>
<td>.de</td>
<td>36940</td>
</tr>
<tr>
<td>.fr</td>
<td>32361</td>
</tr>
<tr>
<td>.au</td>
<td>8577</td>
</tr>
<tr>
<td>.es</td>
<td>5570</td>
</tr>
<tr>
<td>.ca</td>
<td>4717</td>
</tr>
<tr>
<td>.pl</td>
<td>3753</td>
</tr>
<tr>
<td>.it</td>
<td>3405</td>
</tr>
<tr>
<td>.nl</td>
<td>3163</td>
</tr>
<tr>
<td>.uk</td>
<td>3150</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country</th>
<th># Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>.us</td>
<td>238,126</td>
</tr>
<tr>
<td>.fr</td>
<td>183,148</td>
</tr>
<tr>
<td>.de</td>
<td>121,749</td>
</tr>
<tr>
<td>.nl</td>
<td>69,225</td>
</tr>
<tr>
<td>.es</td>
<td>53,548</td>
</tr>
<tr>
<td>.pl</td>
<td>48,041</td>
</tr>
<tr>
<td>.ar</td>
<td>46,265</td>
</tr>
<tr>
<td>.ca</td>
<td>45,523</td>
</tr>
<tr>
<td>.it</td>
<td>40,872</td>
</tr>
<tr>
<td>.no</td>
<td>39,114</td>
</tr>
</tbody>
</table>
Summary/Conclusions
• Service providers don’t make it easy!
• A UUID would help with de-duplication
• Can’t rely on People/Organisations creating metadata
Questions?

OGC Search Engine:  
www.geoseer.net

Contact:  
jonathan.moules@geoseer.net