

Schema.org dataset discovery via dataset Metadata

@pvanguenuchten

GeoCat - Linked Data Factory
Interactive Instruments

Improve discoverability on the web

- Audience outside GEO is interested in GEO data, but is not aware of the GEO standards
- Challenge for a geonetwork is to add support for popular conventions of the web
- Geonetwork can be a facilitator to make WFS data crawlable and downloadable for non GEO users

Metadata Catalog / CSW

- Traditionally supported protocols CSW, OAI-PMH, Z39.50, should be updated to support
 - Restfull API's
 - Crawlability by search engines
- Trational Output formats xml, rdf-xml, html
 - Add Json, json-ld, RDFa, Microdata
- Ontologies such as iso19115, DCAT
 - Add schema.org/Dataset

Challenges

- Crawlability
 - Sitemaps.org spec
 - Crawlable content on persistent url's
- Schema.org mapping
 - Sometimes more detailed sometimes more specific
- How to manage representations on a single url? Should it be a single url?
 - Content negotiation is not enough?

Approach

- Make sure the sitemap and crawlable content url's are set up correctly in GeoNetwork
- Harvest nationaalgeoregister.nl (8000+ datasets)
- Expose sitemap to search engines
- Iterate on crawling issues (web master tools)

Adding microdata to html representation of iso19139

C:\work\geo4web\core-geonetwork\web\src\main\webapp\WEB-INF\data\data\formatter\xslt\render-layout.xml - Notepad++

Bestand Bewerken Zoeken Beeld Coding Syntax Instellingen Macro Uitvoeren Plugins Vensters ?

```

17 <xsl:template mode="fmtfooter" match="undefined"/>
18 <!-- Those templates should be overridden in the schema plugin - end -->
19
20 <!-- Starting point -->
21 <xsl:template match="/">
22   <html>
23     <head><title><xsl:apply-templates mode="getMetadataTitle" select="$metadata"/></title></head>
24     <body>
25       <xsl:apply-templates mode="fmthead" select="$metadata"/>
26       <div class="container gn-metadata-view">
27         <article itemscope="itemscope" itemtype="http://schema.org/Dataset">
28           <header>
29             <h1 itemprop="name"><xsl:apply-templates mode="getMetadataTitle" select="$metadata"/></h1>
30             <!--<p><xsl:apply-templates mode="getMetadataAbstract" select="$metadata"/></p>-->
31             <!-- TODO : Add thumbnail to header -->
32             <!--<xsl:apply-templates mode="render-toc" select="$viewConfig"/>-->
33           </header>
34           <xsl:apply-templates mode="render-view" select="$viewConfig/*"/>
35           <!--
36           TODO: scrollspy or tabs on header ?
37           <div class="gn-scroll-spy"
38             data-gn-scroll-spy="gn-metadata-view-{$metadataId}"
39             data-watch=""
40             data-filter="div > h3"/>-->
41         </article>
42       </div>
43     </body>
44   </html>
45 </template>

```

Add content negotiation

```
124 <rule>
125   <condition name="Accept">application/rdf+xml</condition>
126   <from>^/doc/dataset/(.*)/(.*)</from>
127   <to type="forward">/srv/dut/rdf.metadata.get?uuid=$1</to>
128 </rule>
129
130 <rule>
131   <condition name="Accept">application/json</condition>
132   <from>^/doc/dataset/(.*)/(.*)</from>
133   <to type="forward">/srv/dut/xml.metadata.get?_content_type=json&uuid=$1</to>
134 </rule>
135
136 <rule>
137   <condition name="Accept">text/html</condition>
138   <from>^/doc/dataset/(.*)$</from>
139   <to type="forward">/srv/dut/md.format.xml?xsl=schema-org&uuid=$1</to>
140 </rule>
141
142 <rule>
143   <condition name="Accept">application/xhtml+xml</condition>
144   <from>^/doc/dataset/(.*)$</from>
145   <to type="forward">/srv/dut/md.format.xml?xsl=schema-org&uuid=$1</to>
146 </rule>
147
148 <rule>
149   <from>^/doc/dataset/(.*)$</from>
150   <to type="forward">/srv/dut/xml.metadata.get?uuid=$1</to>
151 </rule>
```

Register sitemap

Search Console

Dashboard

Berichten

► Zoekopmaak ⓘ

► Zoekverkeer

► Google-index

▼ Crawlen

Crawlfouten

Crawlstatistieken

Fetchen als Google robots.txt-tester

Sitemaps

URL-parameters

Beveiligingsproblemen

Andere bronnen

Sitemapindex

Index: [/srv/dut/portal.sitemap](#)

Deze Sitemapindex is verzonden op 7 dec. 2015 en verwerkt op 7 dec. 2015.

Sitemaps in deze index Sitemapfouten Indexfouten

Sitemapinhoud

Alle inhoudstypen

■ Verzonden

Webpagina's

1.661 Verzonden

2.000

1.500

1.000

500

Web

Sitemaps in deze sitemapindex (Alle inhoudstypen)

Weergeven 25 rijen 1-7 van 7

#	Sitemap ▲	Type	Verwerkt	Problemen	Items	Verzonden	Geïndexeerd
1	/sitemap/1/dut	Sitemap	7 dec. 2015	-	Web	250	-
2	/sitemap/2/dut	Sitemap	7 dec. 2015	-	Web	250	-
3	/sitemap/3/dut	Sitemap	7 dec. 2015	-	Web	250	-
4	/sitemap/4/dut	Sitemap	7 dec. 2015	-	Web	250	-
5	/sitemap/5/dut	Sitemap	7 dec. 2015	-	Web	250	-
6	/sitemap/6/dut	Sitemap	7 dec. 2015	-	Web	250	-
7	/sitemap/7/dut	Sitemap	7 dec. 2015	-	Web	161	-

1-7 van 7

Review crawl errors

Search Console

Dashboard

Berichten

Zoekopmaak

Zoekverkeer

Google-index

Crawlen

Crawlfouten

Crawlstatistieken

Fetchen als Google

robots.txt-tester

Sitemaps

URL-parameters

Beveiligingsproblemen

Andere bronnen

Sitefouten

Geen fouten gevonden in de afgelopen 90 dagen. Goed gedaan!

URL-fouten

Status: 7-12-15

Desktop Smartphone

Serverfout Niet gevonden

5 74

10,0

7,5

5,0

2,5

4-12-15 5-12-15 6-12-15 7-12-15

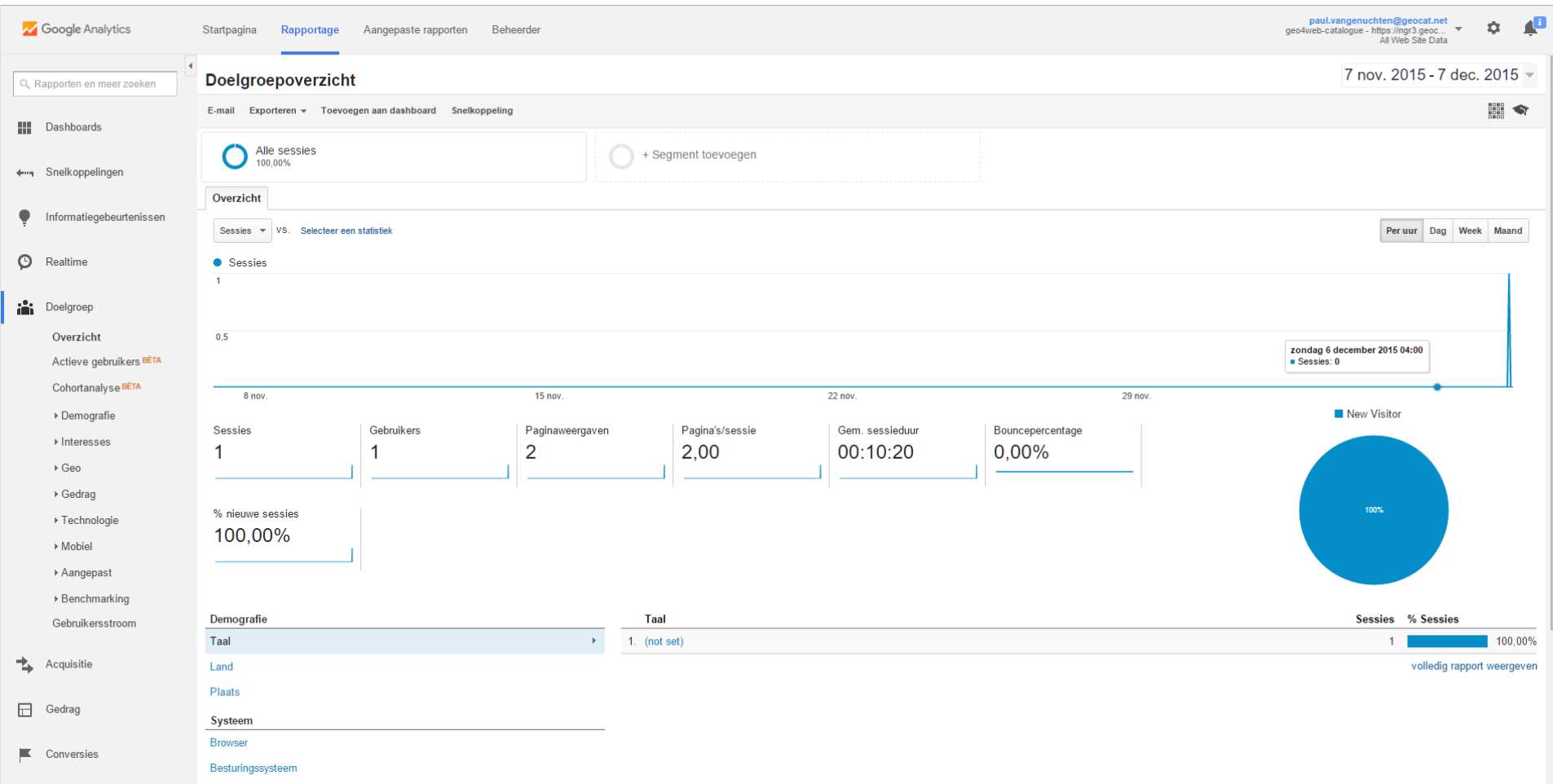
Top 1000 pagina's met fouten

Downloaden MARKEREN ALS GECORRIGEERD (0) Filter

Weergeven 25 rijen 1-5 van 5

Prioriteit	URL	Reactiecode	Gedetecteerd
1	srv/eng/xml.metadata.get?id=839	500	5-12-15
2	doc/dataset/B8F8D2A2-F599-4262-B4C4-63239FA0E952	503	4-12-15
3	doc/dataset/82120933-3648-464F-8B74-1976434460B1	503	4-12-15
4	doc/dataset/366244B5-79E2-4D8C-983F-8521807EEE6A	503	4-12-15
5	doc/dataset/d335c66d-1823-43c5-bf05-0c079daf51f6	503	4-12-15

Analyse usage statistics



Structured data testing tool



Fetch URL

Examples ▾

VALIDATE

Results - Filter by use case ▾

Dataset (1)

16 Errors

Dataset:	http://www.example.com/gn-metadata-view-2526
name:	Beneden Merwede vaargeul en oeverloding grid opn.09-2005
thumbnailUrl:	http://geoservices.rws.nl/opendata/img/BEM006.jpg
keywords:	HoogteVervoersnetwerken
keywords:	aardrijkskundehoogtelandschapwaterBeneden Merwede vaargeul en oeverlodingBEM006
keywords:	hoogte
fileFormat:	()
isBasedOnUrl:	http://www.example.com/xml.metadata.get?id=2526
author [Organization]:	
email:	mailto:servicedesk-data@rws.nl
address [PostalAddress]:	
streetAddress:	Derde Werelddreef 1
addressLocality:	Delft
addressRegion:	Zuid-Holland
postalCode:	2622 HA
addressCountry [Country]:	
name:	Nederland
contactPoint [ContactPoint]:	

```
34     </div>
35   </div>
36 </nav>
37 <div class="container gn-metadata-view">
38   <article id="gn-metadata-view-2526" itemscope="itemscope"
itemtype="http://schema.org/Dataset">
39     <header>
40       <h1 itemprop="name">Beneden Merwede vaargeul en oeverloding grid
opn.09-2005</h1>
41     </header>
42     <div id="gn-tab-default">
43       <h3 class="view-header">Standaard</h3>
44       <div id="gn-view-d276e1592">
45         <dl>
46           <dt>Titel</dt>
47           <dd>Beneden Merwede vaargeul en oeverloding grid opn.09-2005</dd>
48         </dl>
49         <dl>
50           <dt>Alternatieve titel</dt>
51           <dd>Projectcode: BEM006</dd>
52         </dl>
53         <dl class="gn-date">
54           <dt>
55             Datum
56           </dt>
57           <span title="Datum waarop de dataset of dataset serie is
gecreïerd.">creatie</span>
```

WFS proxy

- Metadata in iso19139 typically references a WFS endpoint
- Metadata in Schema.org should reference a WFS proxy (currently in development by interactive instruments) that grabs data from the WFS and exposes in schema.org (as RDFa and/or json-ld)

Catalog as a proxy

- A wfs-getcapabilities typically refers to an item in a catalog, but that content may not (yet) be available in the catalog linked to the wfs proxy
 - Catalog should grab the content from the designated catalog and transform to schema.org on the fly