Schema.org dataset discovery via dataset Metadata

@pvangenuchten

GeoCat - Linked Data Factory

Interactive Instruments

Improve discoverability on the web

- Audiance 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.xsl - Notepad++

Bestand Bewerken Zoeken Beeld Codering Syntaxis Instellingen Macro Uitvoeren Plugins Vensters ?

urlrewrite.xml | render-layout.xsl | view.xsl | snippets.txt | Sitemap.java | sitemap.xsl | view.xsl | render-variables.xsl | NL_profiel_op_ISO19115_v13_alleen_verplicht.xml

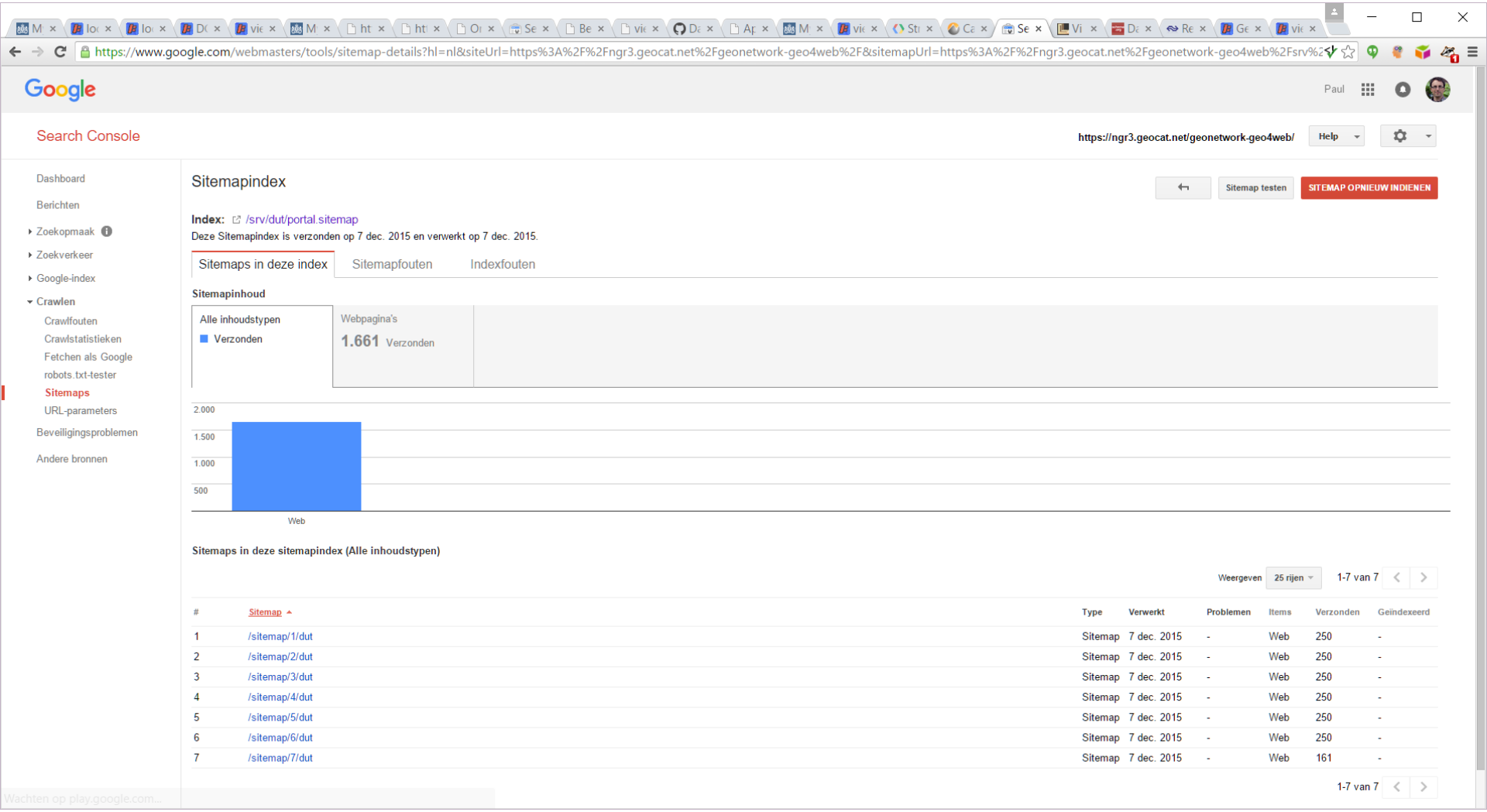
```
17    <xsl:template mode="fmtfooter" match="undefined"/>
18    <!-- Those templates should be overriden in the schema plugin - end -->
19
20    <!-- Starting point -->
21    <xsl:template match="/">
22      <html>
23      <head><title><xsl:apply-templates mode="getMetadataTitle" select="$metadata"/></title></head>
24      <body>
25      <xsl:apply-templates mode="fmtheader" select="$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          <footer>
42
```

Add content negotiation

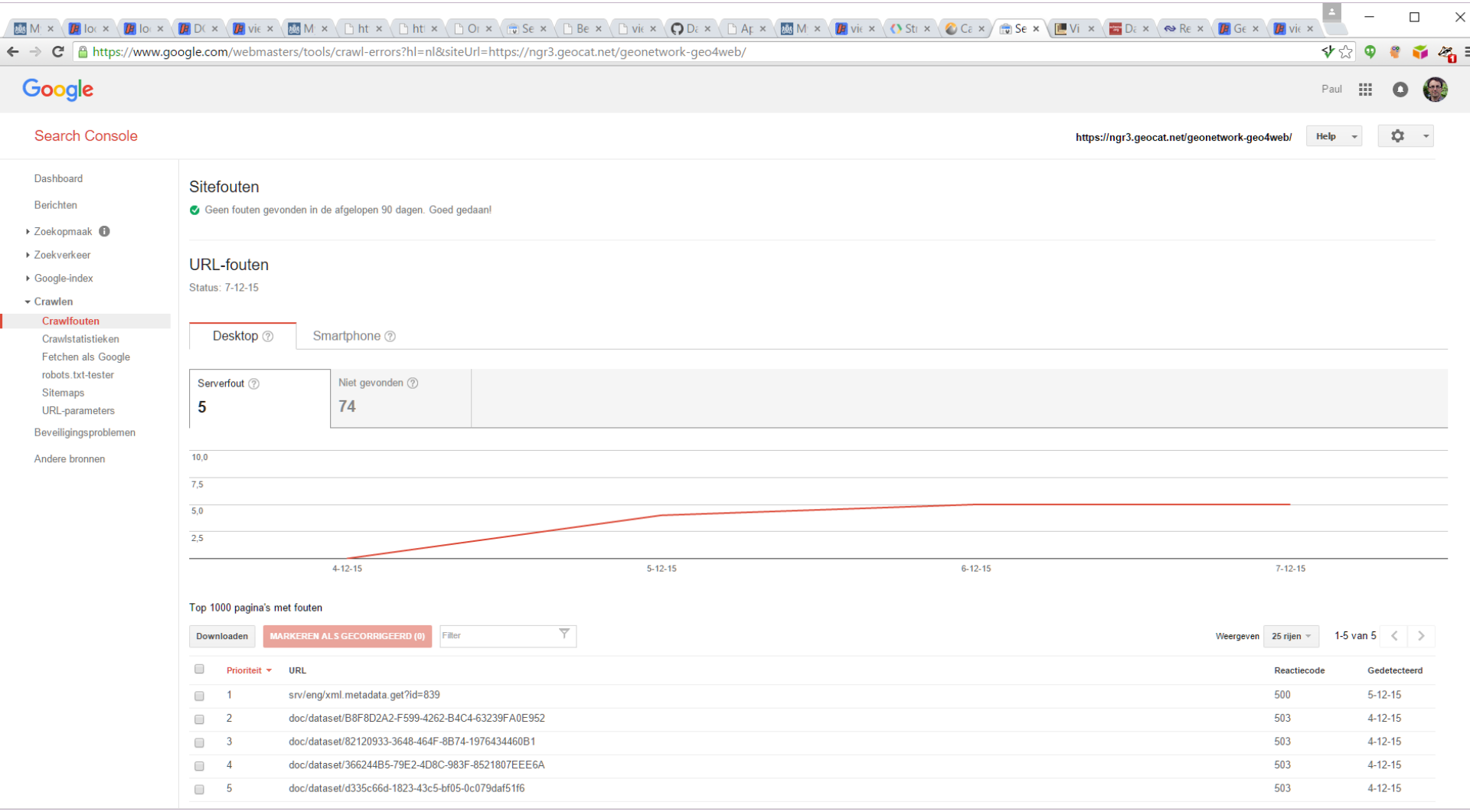
```
124     <rule>
125         <condition name="Accept">application/rdf\+xml</condition>
126         <from>^/doc/dataset/(.*)/(.*)</from>
127         <to type="forward">/srv/dut/rdf.metadata.get?uuid=$1</to>
128     </rule>
129
130     <rule>
131         <condition name="Accept">application/json</condition>
132         <from>^/doc/dataset/(.*)/(.*)</from>
133         <to type="forward">/srv/dut/xml.metadata.get?_content_type=json&amp;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&amp;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&amp;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>
```

eXtensible Markup Language file | length : 12708 lines : 368 | Ln : 123 Col : 1 Se

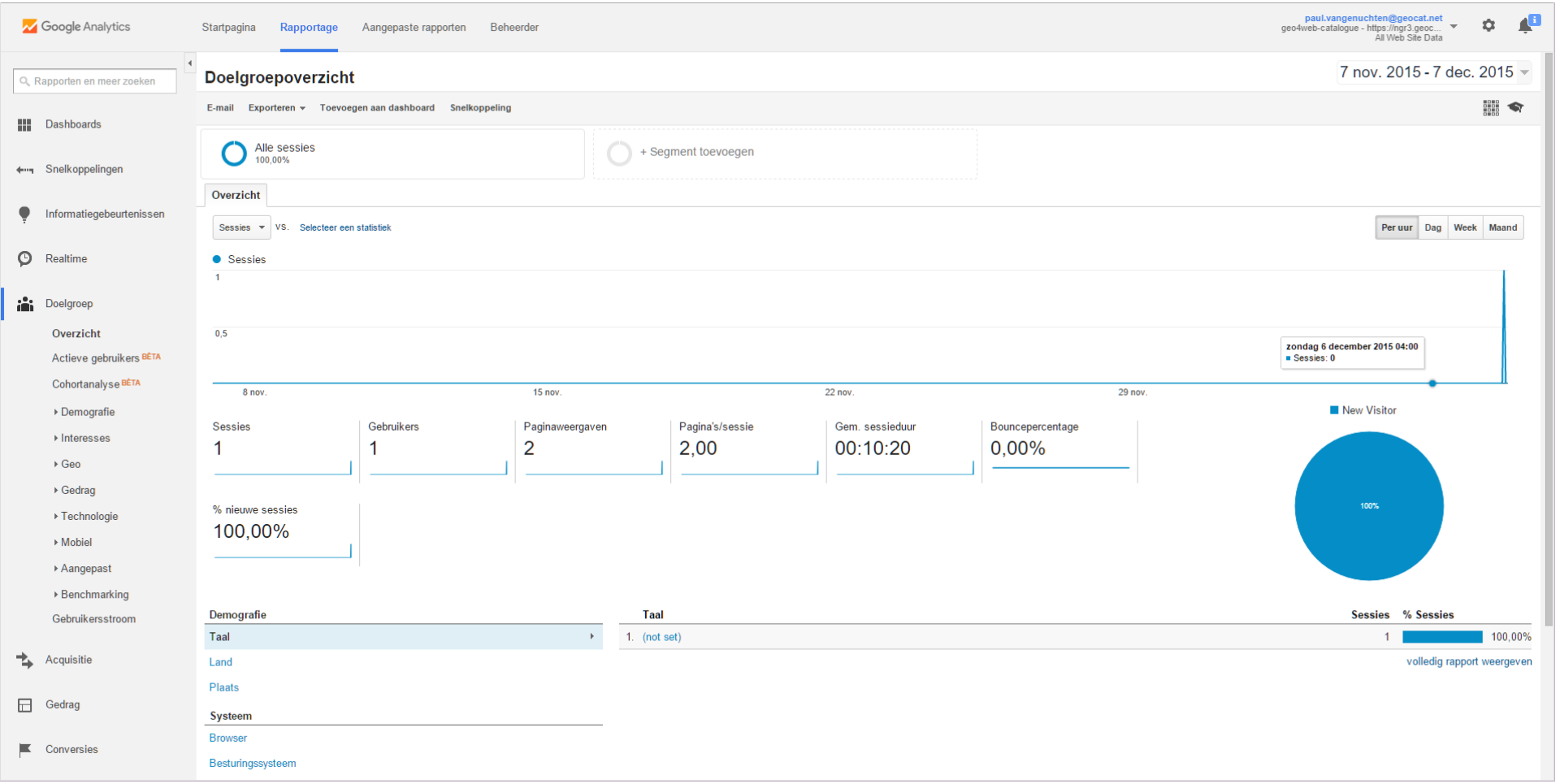
Register sitemap



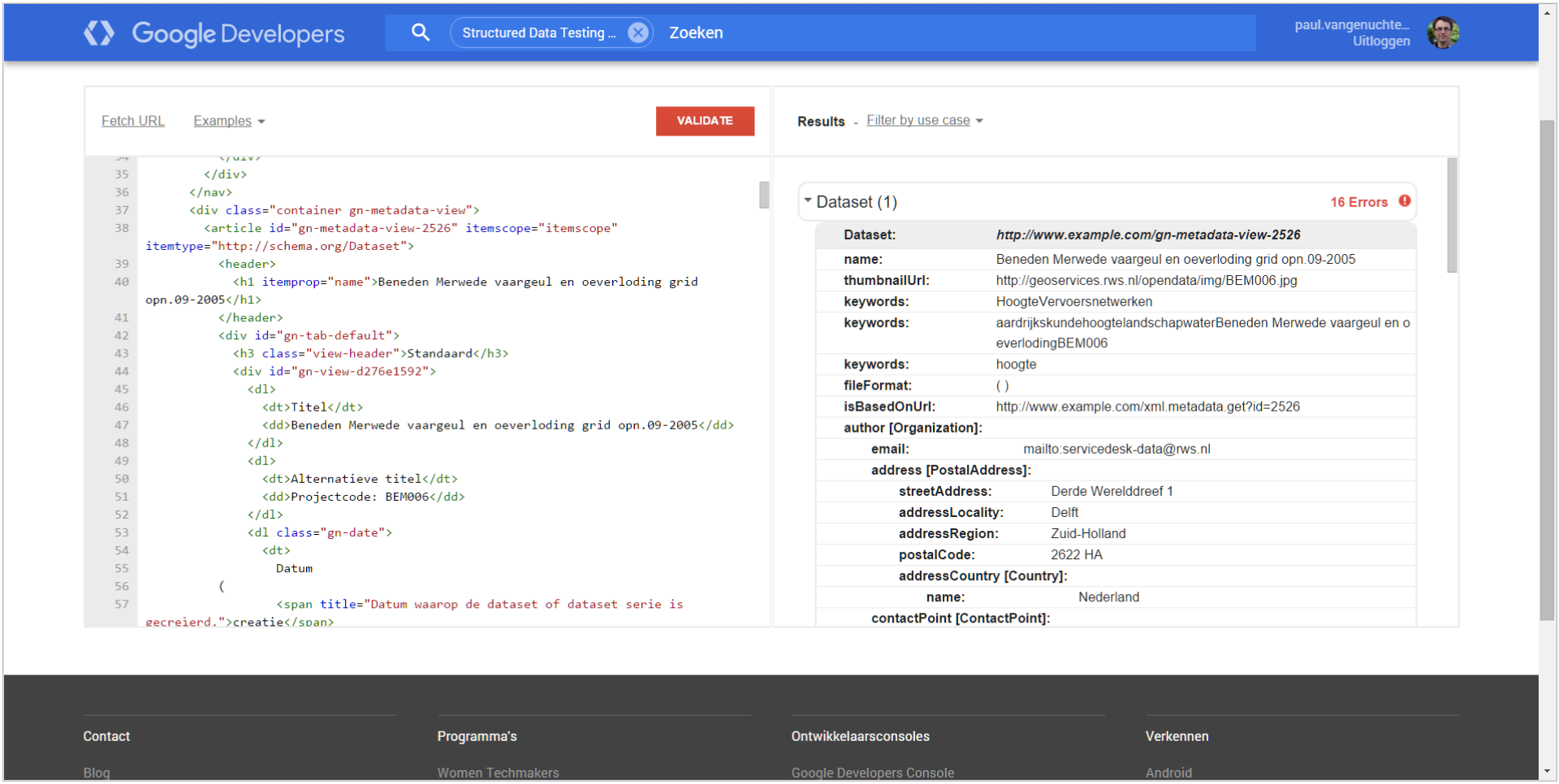
Review crawl errors



Analyse usage statistics



Structured data testing tool



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 transorm to schema.org on the fly