GeoNetwork as a facilitator of Search Engine Discoverability of iso19115 records

INSPIRE Discovery Workshop 2019

Paul van Genuchten

Linda van de Brink

Ispra, 3-4 july 2019

Contents

These days GeoNetwork records are still hardly harvested by search engines. GeoNetwork typically presents iso19115 data in a html format, potentially ideal for search engine ingestion. However GeoNetwork by design has some challenges that prevent content to be easily ingested. In this presentation I’ll present some optimizations that have been introduced in recent GeoNetwork versions related to search engine ingestion and some experiences with search engine optimization in GeoNetwork from the GeoNovum testbed Geo4Web.
Optimisations focus on the use of Search Engine Console monitoring, use of a sitemap, URI strategy, schema.org annotations and indicating which parts of the catalog not to crawl.

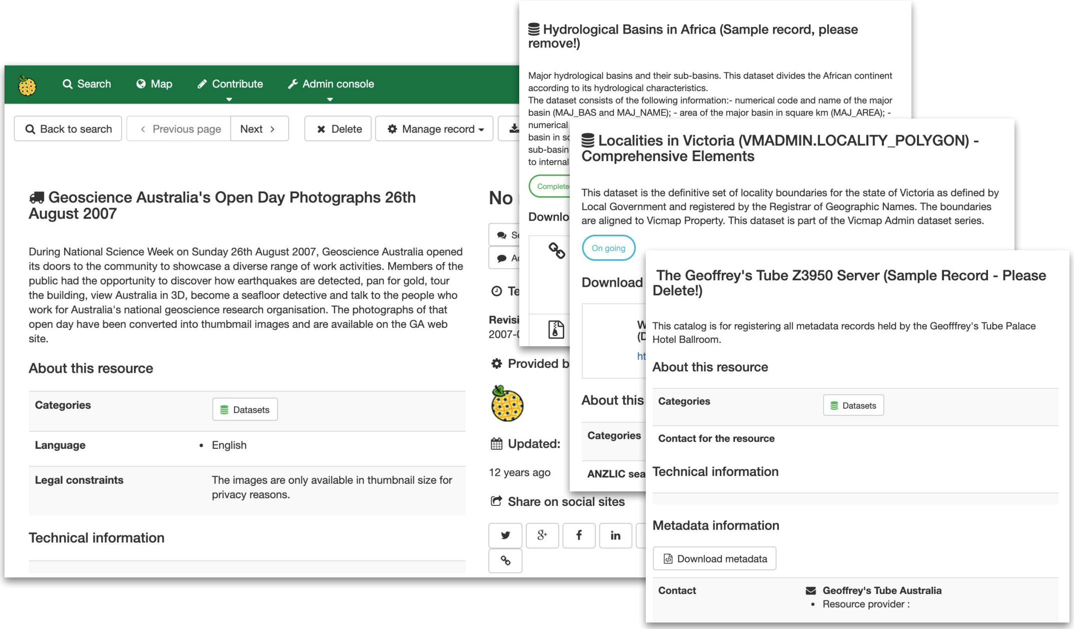
GeoNetwork Opensource

Catalogue software to register spatial assets, datasets & services.

Allows multiple metadata schema’s, mainly ISO19115 and ISO19115-2

Facilitates ISO to html conversion, but not commonly indexed in search engines.

Search engines 🚫♥ JS frameworks



Search | Map | Contribute | Admin console

Back to search | ‹ Previous page | Next › | ✖ Delete | ⚙ Manage record

Geoscience Australia's Open Day Photographs 26th August 2007

During National Science Week on Sunday 26th August 2007, Geoscience Australia opened its doors to the community to showcase a diverse range of work activities. Members of the public had the opportunity to discover how earthquakes are detected, pan for gold, tour the building, view Australia in 3D, become a seafloor detective and talk to the people who work for Australia's national geoscience research organisation. The photographs of that open day have been converted into thumbnail images and are available on the GA web site.

About this resource

Categories	Datasets
Language	- English
Legal constraints	The images are only available in thumbnail size for privacy reasons.

Technical information

No

Downlo

Te

Revisi
2007-(

Provided b

Updated:

12 years ago

Share on social sites

Hydrological Basins in Africa (Sample record, please remove!)

Major hydrological basins and their sub-basins. This dataset divides the African continent according to its hydrological characteristics.
The dataset consists of the following information:- numerical code and name of the major basin (MAJ_BAS and MAJ_NAME); - area of the major basin in square km (MAJ_AREA); - numerical
basin in sc
sub-basin
to internal

Complete

Downlo

Localities in Victoria (VMADMIN.LOCALITY_POLYGON) - Comprehensive Elements

This dataset is the definitive set of locality boundaries for the state of Victoria as defined by Local Government and registered by the Registrar of Geographic Names. The boundaries are aligned to Vicmap Property. This dataset is part of the Vicmap Admin dataset series.

On going

Download

About this

Categories

ANZLIC sea

The Geoffrey's Tube Z3950 Server (Sample Record - Please Delete!)

This catalog is for registering all metadata records held by the Geofffrey's Tube Palace Hotel Ballroom.

About this resource

Categories	Datasets
Contact for the resource	

Technical information

Metadata information

Download metadata

Contact	Geoffrey's Tube Australia - Resource provider :

=1

Search engine requires single URL for each metadata

An alternative UI

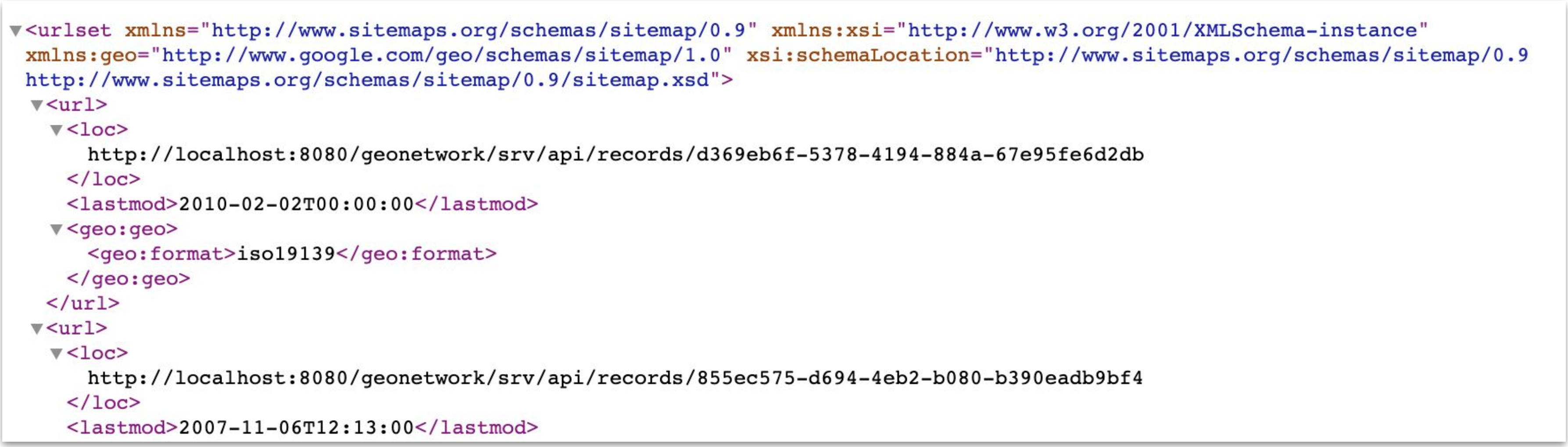
Geonetwork traditionally has a no-js interface for non-js-aware clients.

This interface is crawled by search engines

A button to return to the default layout

Sitemap.xml

Search engines allow to register a sitemap, to increase crawling speed.



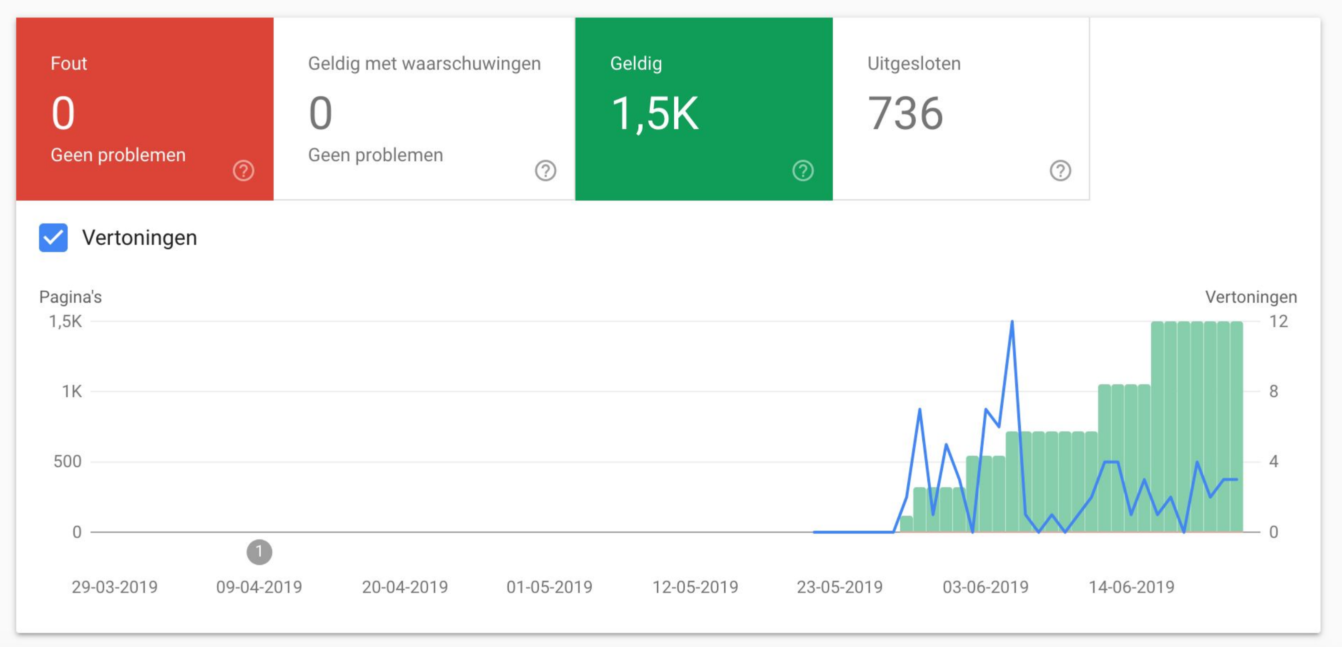
```
▼<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns:geo="http://www.google.com/geo/schemas/sitemap/1.0" xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9
 http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd">
  ▼<url>
    ▼<loc>
        http://localhost:8080/geonetwork/srv/api/records/d369eb6f-5378-4194-884a-67e95fe6d2db
      </loc>
      <lastmod>2010-02-02T00:00:00</lastmod>
    ▼<geo:geo>
        <geo:format>iso19139</geo:format>
      </geo:geo>
    </url>
  ▼<url>
    ▼<loc>
        http://localhost:8080/geonetwork/srv/api/records/855ec575-d694-4eb2-b080-b390eadb9bf4
      </loc>
      <lastmod>2007-11-06T12:13:00</lastmod>
```

Robots.txt

A ‘configuration’ file to tell the browser how to crawl the website.

- Location of sitemap.xml
- Which paths to exclude (xml, json)

Web console



Fout

0

Geen problemen

Geldig met waarschuwingen

0

Geen problemen

Geldig

1,5K

Uitgesloten

736

☑ Vertoningen

Register and evaluate the search/crawl behaviour

Additional benefits

Find out what search terms people use

Find broken links in metadata

Schema.org/Dataset

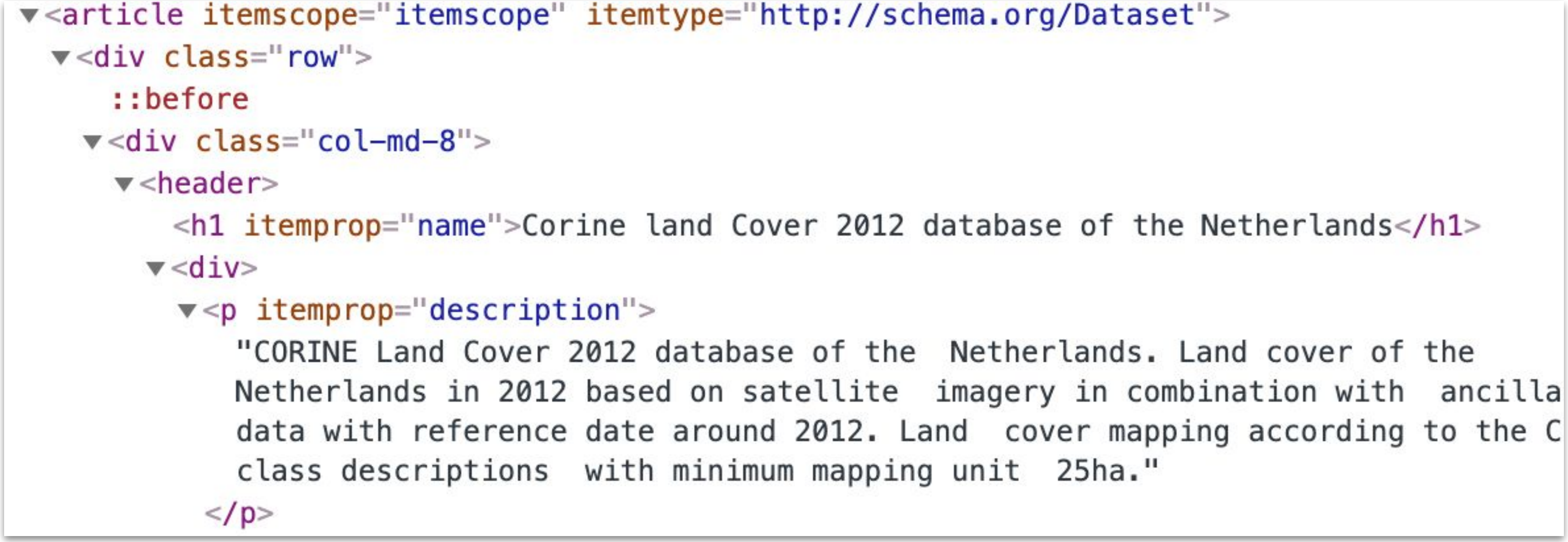
Search engines use the schema.org ontology to crawl content in a structured way

Schema.org/Dataset describes datasets

Google launched a dataset search engine based on this structured data.

Schema.org in GeoNetwork

First approach used microdata

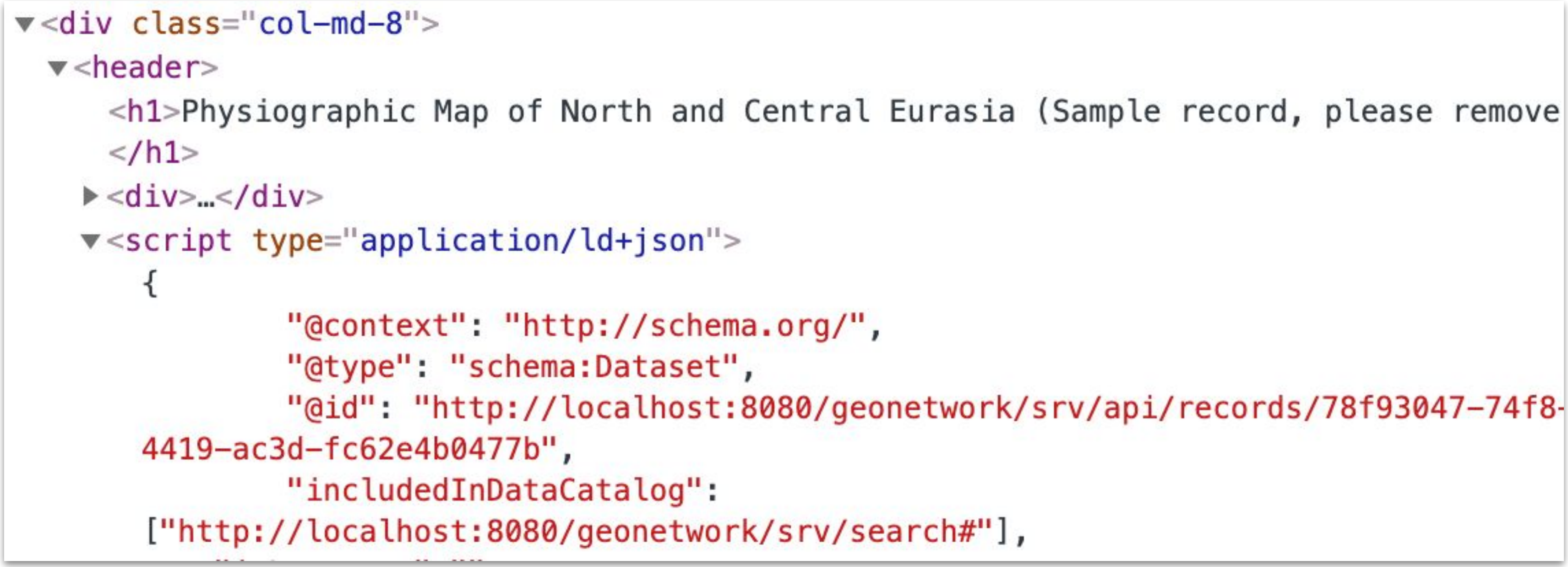


```
▼<article itemscope="itemscope" itemtype="http://schema.org/Dataset">
  ▼<div class="row">
      ::before
    ▼<div class="col-md-8">
      ▼<header>
          <h1 itemprop="name">Corine land Cover 2012 database of the Netherlands</h1>
        ▼<div>
          ▼<p itemprop="description">
              "CORINE Land Cover 2012 database of the  Netherlands. Land cover of the
              Netherlands in 2012 based on satellite  imagery in combination with  ancilla
              data with reference date around 2012. Land  cover mapping according to the C
              class descriptions  with minimum mapping unit  25ha."
            </p>
```

Embedded Json-ld

In 3.8 we are switching to embedded json-ld

- Support multiple schema’s
- Metadata also available as plain json-ld triples
- Cleaner approach
- Json-ld is now available in multiple search engine crawlers



```
▼<div class="col-md-8">
  ▼<header>
      <h1>Physiographic Map of North and Central Eurasia (Sample record, please remove
      </h1>
    ▶<div>…</div>
    ▼<script type="application/ld+json">
        {
                "@context": "http://schema.org/",
                "@type": "schema:Dataset",
                "@id": "http://localhost:8080/geonetwork/srv/api/records/78f93047-74f8-
        4419-ac3d-fc62e4b0477b",
                "includedInDataCatalog":
        ["http://localhost:8080/geonetwork/srv/search#"],
```

Conclusions

GeoNetwork has a role in making resources discoverable via search engines

Search engine console is a useful tool to improve catalogue contents

JS frameworks not optimal, but there are workarounds

Thank you!