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
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 do not love JS frameworks

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.

```xml
  <url>
    <loc>http://localhost:8080/geonetwork/srv/api/records/d369eb6f-5378-4194-884a-67e95fe6d2db</loc>
    <lastmod>2010-02-02T00:00:00</lastmod>
  </url>
  <url>
    <loc>http://localhost:8080/geonetwork/srv/api/records/855ec575-d694-4eb2-b080-b390eadb9bf4</loc>
    <lastmod>2007-11-06T12:13:00</lastmod>
  </url>
</urlset>
```
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

Register and evaluate the search/crawl behaviour
Additional benefits

Find out what search terms people use
Find broken links in metadata
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

```html
<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>
      </header>
      <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>
      </div>
    </div>
  </div>
</article>
```
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>
</header>
<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#"]
}
</script>
</div>
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!