Facilitating FAIR Standards through Google Page Map Implementation at the Canadian Cryospheric Information Network/Polar Data Catalogue

Third Polar Data Forum
November 19, 2019
Helsinki, Finland
Gregory Vey & Wesley Van Wychen
Overview

• Findability: A FAIR Principle
• Search Engine Optimization
• Google PageMap
• PageMap Implementation
• Limitations
• Conclusion
• Future Work
Findability: A FAIR Principle

• To Be Findable:

• F1. (meta)data are assigned a globally unique and eternally persistent identifier.
• F2. data are described with rich metadata.
• F3. (meta)data are registered or indexed in a searchable resource.
• F4. metadata specify the data identifier.
Search Engine Optimization (SEO)

• **General Perspective**
• The process of increasing the visibility of a website to users of a web search engine
• For the purpose of increasing the quality and quantity of website traffic
• Involves editing or adding content, and modifying HTML and associated coding to increase relevance to specific keywords and remove barriers to the indexing activities of search engines
• Google Perspective
  • Help Google find your content
  • Tell Google which pages shouldn't be crawled
  • Help Google (and users) understand your content
  • Manage your appearance in Google Search results
  • Organize your site hierarchy
  • Optimize your content and images
  • Make your site mobile-friendly
  • Promote your website
  • Analyze your search performance and user behavior
SEO at PDC
SEO at PDC

Number of Metadata: 2,969
Number of Datasets: 437
Number of Datafiles: >2,900,000 (~21 TB)
Number of DOIs assigned: 375
Number of RADARSAT Files: >27,000 (8.5 TB)
Total Size for all Archived Data: 37 TB

Dynamic URLs
Google PageMap

• **SiteMap:**
  • From an SEO perspective, a structured XML format that lists the pages in a site, their relative importance, and how often they are updated
  • The Sitemaps protocol allows publication of links because many sites have dynamic pages available through forms and user input: A Sitemap contains URLs to these pages so that web crawlers can find them

• **PageMap:**
  • A structured XML format created by Google to enable website creators to embed data and notes in their webpages
  • A separate schema intended to be embedded within a SiteMap URL Element
Google PageMap

• SiteMap versus PageMap

SiteMap

PageMap

```xml
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.example.net/?id=who</loc>
    <lastmod>2009-09-22</lastmod>
    <changelang>monthly</changelang>
    <priority>0.8</priority>
  </url>
  <url>
    <loc>http://www.example.net/?id=what</loc>
    <lastmod>2009-09-22</lastmod>
    <changelang>monthly</changelang>
    <priority>0.5</priority>
  </url>
  <url>
    <loc>http://www.example.net/?id=how</loc>
    <lastmod>2009-09-22</lastmod>
    <changelang>monthly</changelang>
    <priority>0.5</priority>
  </url>
</urlset>
```

```
<xml version="1.0" encoding="UTF-8">
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.example.net/?id=who</loc>
    <lastmod>2009-09-22</lastmod>
    <changelang>monthly</changelang>
    <priority>0.8</priority>
  </url>
  <PageMap xmlns="http://www.google.com/schemas/sitemap-pagemap/1.0">
    <DataObject type="document" id="bibachi">
      <Attribute name="name">Dragon</Attribute>
      <Attribute name="review">3.5</Attribute>
    </DataObject>
  </PageMap>
</urlset>
```

SiteMap versus PageMap

SiteMap

PageMap
**Google PageMap**

- **PageMap Tags:**

<table>
<thead>
<tr>
<th>Tag</th>
<th>Required</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageMap</td>
<td>Yes</td>
<td>Encloses all PageMap information for the relevant URL</td>
</tr>
<tr>
<td>DataObject</td>
<td>Yes</td>
<td>Encloses all information about a single element (for example, an action) that should appear in the Custom Search result</td>
</tr>
<tr>
<td>Attribute</td>
<td>Yes</td>
<td>Each DataObject contains one or more attributes</td>
</tr>
</tbody>
</table>
PageMap Implementation

- **Acquire Schemas:**
  - SiteMap XSD ([http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd](http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd))
  - PageMap XSD ([http://www.google.com/schemas/sitemap-pagemap/1.0/sitemap-pagemap.xsd](http://www.google.com/schemas/sitemap-pagemap/1.0/sitemap-pagemap.xsd))

- **Build Standard XML Marshaller Project**
  - PDC uses JAXB
  - Input Logic
    - Metadata SQL Query *(this will be specific to the given repo)*
  - Generate SiteMap with embedded PageMaps *(i.e. Marshal Metadata)*

- **Expose SiteMap for Crawlers**
PageMap Implementation

Project Structure

- ccadipagemap [ccadip-page-map master]
  - JRE System Library [JavaSE-1.8]
  - src
    - (default package)
      - com.pdc.ccadixml
        - MetadataReader.java
        - PdcUnmarshaller.java
        - RecordFactory.java
      - com.pdc.ccadixmlresources
        - sitemap-pagemap.xsd
        - sitemap.xsd
      - com.pdc.ccadixml.sitemap
        - ObjectFactory.java
        - TChangeFreq.java
        - TLM.java
        - UrlSet.java
      - com.pdc.ccadixml.sitemap.pagemap
        - ObjectFactory.java
        - PageMap.java
        - log4j.properties
      - Maven Dependencies
      - EclipseLink 2.5.2
      - bin
      - log
      - target
      - pom.xml
PageMap Implementation

Sample PageMap (embedded in SiteMap)

Submit full SiteMap for indexing via the Google Search Console
Limitations

• To Be Interoperable:

  • I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  • I2. (meta)data use vocabularies that follow FAIR principles.
  • I3. (meta)data include qualified references to other (meta)data.
Limitations
Limitations

Currently not recognized by Testing Tool

1. metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation

CCIN
Canadian Cryospheric Information Network

Polardata Catalogue
Limitations

• Full SEO Requires Other Formats:

```
  "site":{
    "@type":"Dataset",
    "name":"ERA5 global reanalysis precipitation, evaporation, and runoff for 2002-2017",
    "description":"ERA5 is the fifth generation atmospheric reanalysis dataset from European Centre for Medium-Range Weather Forecasts",
    "url":"https://www.polardata.ca/docsearch/PDOCSearch.jsp?did=116875",
    "keywords":
      [
        "Geographic Locations -> Global",
        "Natural sciences -> Atmosphere",
        "Natural sciences -> Clouds",
        "Natural sciences -> Modelling",
        "Natural sciences -> Monitoring",
        "Natural sciences -> Precipitations",
        "Natural sciences -> Surface water runoff",
        "Natural sciences -> Temperature",
        "Natural sciences -> Water"
      ],
    "creator":{
      "@type":"Organization",
      "url":"https://cdn.climate.copernicus.eu/op/search?text=ERA5&type=dataset",
      "name":"Global Water Futures, University of Waterloo",
      "contactPoint":{
        "@type":"ContactPoint",
        "contactType":"Technical support",
        "telephone":"+2510881167 ext 33327",
        "email":"gwf@uwaterloo.ca"
      }
    }
  },
  "includedInDataCatalog":{
    "@type":"DataCatalog",
    "name":"PolarData.ca"
  },
  "temporalCoverage":2002-2017,
  "spatialCoverage":{
    "@type":"Place",
    "geo":{
      "@type":"GeoShape",
      "bbox":50.0 -180.0 -90.0 180.0"
    }
  }
```

11. metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation
Limitations

• Full SEO Requires Other Formats:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERA5 global reanalysis precipitation, evaporation, and runoff for 2003-2017</td>
<td>ERA5 is the fifth generation atmospheric reanalysis dataset from European Centre for Medium-Range Weather Forecasts (ECMWF) that is available to support atmospheric and hydrological research across the globe. This data set will be used to support the modeling objectives of the Global Water Futures (GWF) Program funded by Canadian Arctic Research Excellence Fund.</td>
</tr>
</tbody>
</table>

Testing Tool Result
Conclusion

• **Separate PageMap and JSON-LD Generation Processes:**
  
  • Create SiteMap of dynamic URLs with embedded PageMaps
  
  • Expose SiteMap to crawlers
  
  • Provide dynamic injection of JSON-LD schema.org tags for URLs
Future Work

• Metadata JSON-LD Generation Service:
Thank-you

• Questions?

www.polardata.ca