Experimental use of Sitemaps.org and Schema.org metadata for Search Engine Optimization

by Mark Servilla
Environmental Data Initiative

Trail Map

- Motivation
- What is Search Engine Optimization?
- What is sitemaps.org and schema.org?
- How does EDI use sitemaps.org and schema.org metadata?
- How does SEO affect data stored in the EDI Data Repository?
- Are there other initiatives using sitemaps.org and schema.org?

Finding the "right" environmental data on the Interet is still difficult!

What is Search Engine Optimization?

Search engine optimization (SEO) is the process of affecting the online visibility of a web page in a search engine's unpaid results - often referred to as "natural", "organic", or "earned" results.

- https://en.wikipedia.org/wiki/Search_engine_optimization

What is Search Engine Optimization?

Search engine optimization (SEO) is the process of affecting the online visibility of a web page in a search engine's unpaid results - often referred to as "natural", "organic", or "earned" results.

- https://en.wikipedia.org/wiki/Search_engine_optimization

SEO benefits both web users and web sites by ranking web pages with specific characteristics higher in search results - the most relevant content floats to the top of the ranking so users can more easily find the "good" information

What is Search Engine Optimization?

This results in a more satisfying user experience (web searching) and greater web site exposure - a win-win for all

So how do we implement search engine optimization for environmental data?

We tell search engines what web pages to index and we tell them the details about what to index!

Sitemaps.org and Schema.org Metadata

Sitemaps.org and Schema.org are "markup" protocols (metadata) for web-pages to assist search engines in indexing web-page content:

Sitemaps.org and Schema.org Metadata

Sitemaps.org and Schema.org are "markup" protocols (metadata) for web-pages to assist search engines in indexing web-page content:

- sitemaps.org - a table-of-contents (of sorts) to tell search engines what web pages to index

Sitemaps.org and Schema.org Metadata

Sitemaps.org and Schema.org are "markup" protocols (metadata) for web-pages to assist search engines in indexing web-page content:

- sitemaps.org - a table-of-contents (of sorts) to tell search engines what web pages to index
- schema.org - a "markup" vocabulary to tell search engines standard and detailed information about the data set*

Sitemaps.org and Schema.org Metadata

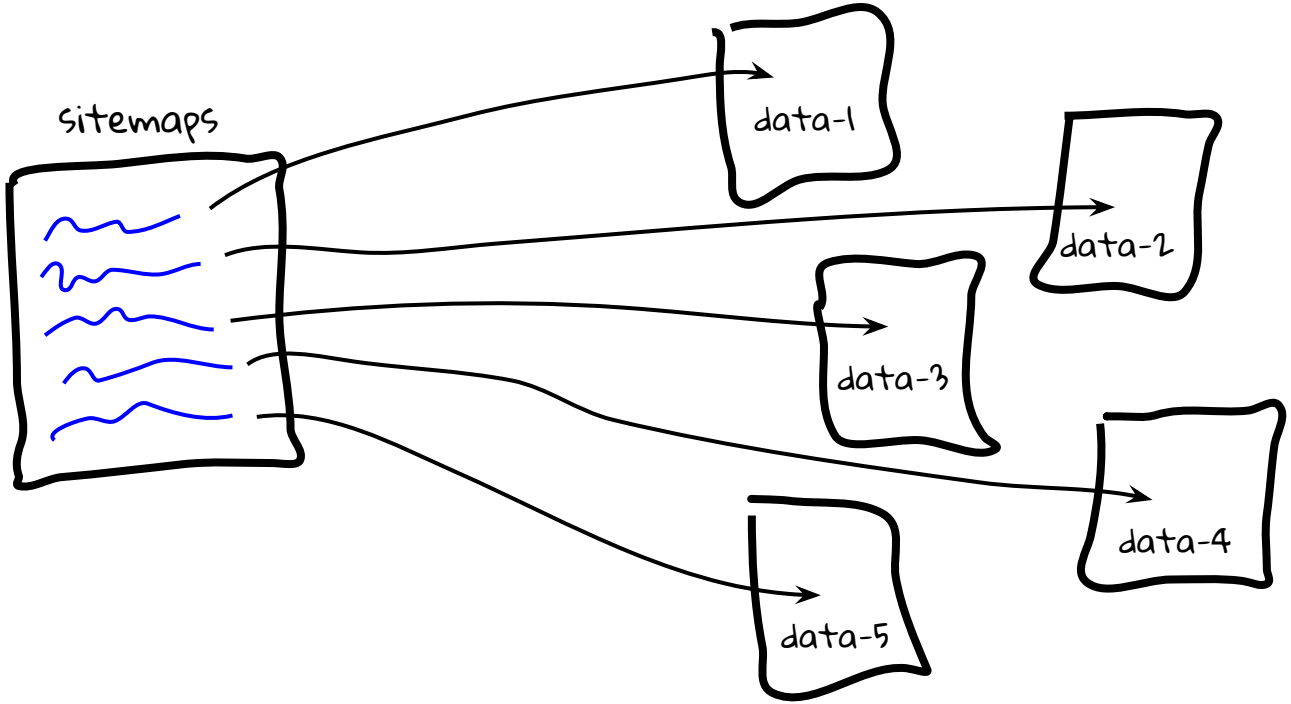
Sitemaps.org and Schema.org are "markup" protocols (metadata) for web-pages to assist search engines in indexing web-page content:

- sitemaps.org - a table-of-contents (of sorts) to tell search engines what web pages to index
- schema.org - a "markup" vocabulary to tell search engines standard and detailed information about the data set*

*schema.org metadata provides a "markup" vocabulary for all types of information found on the Internet, not just environmental data sets

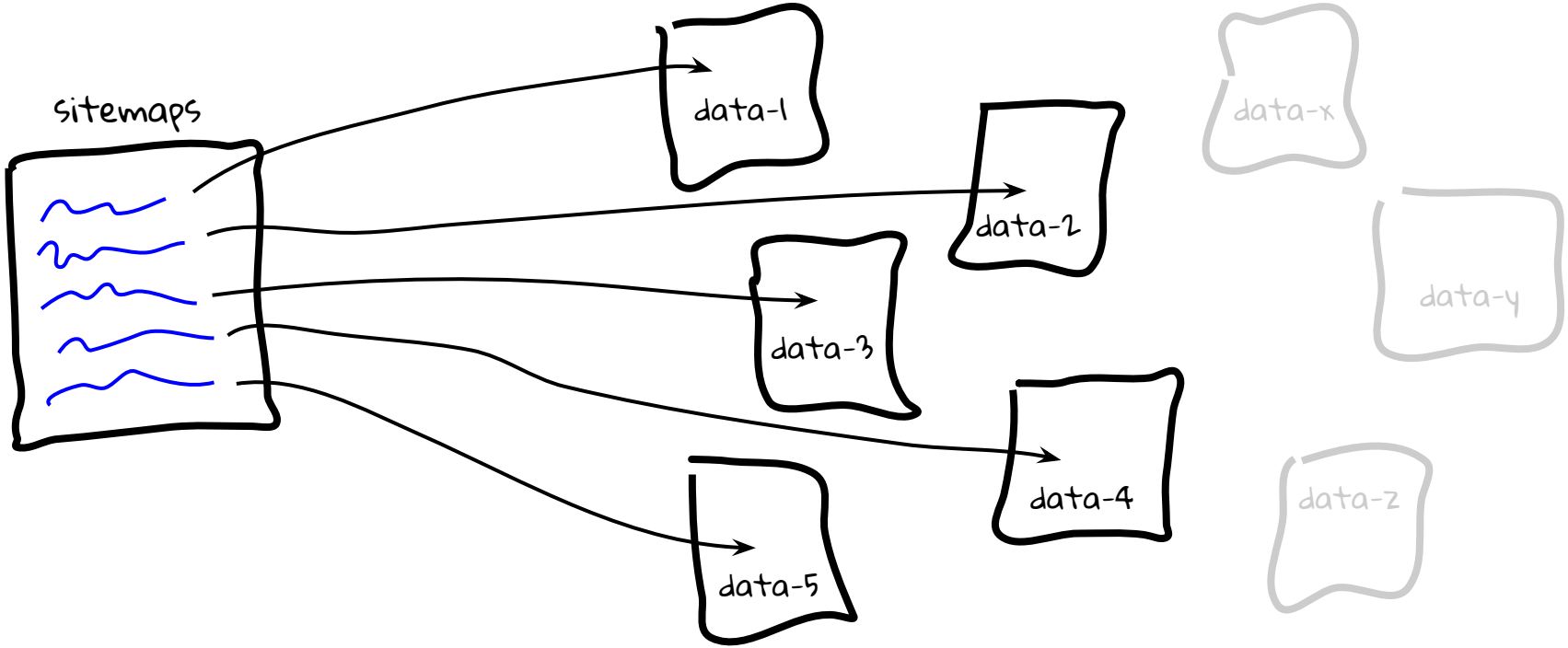
Sitemaps.org Metadata

Provides direct URL links to web-pages that "should" be indexed



Sitemaps.org Metadata

Provides direct URL links to web-pages that "should" be indexed



Sitemaps.org Metadata

- Simple XML protocol

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
   <url>
      <loc>http://www.example.com/</loc>
      <lastmod>2005-01-01</lastmod>
      <changefreq>monthly</changefreq>
      <priority>0.8</priority>
   </url>
</urlset>
```

Sitemaps.org Metadata

- Simple XML protocol

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
   <url>
      <loc>http://www.example.com/</loc>
      <lastmod>2005-01-01</lastmod>
      <changefreq>monthly</changefreq>
      <priority>0.8</priority>
   </url>
</urlset>
```

- Limited number of links (50k) and volume (50MB) of sitemap files
  - Can be extended using a sitemaps index file - TOC of sitemaps files

Sitemaps.org Metadata

- Simple XML protocol

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
   <url>
      <loc>http://www.example.com/</loc>
      <lastmod>2005-01-01</lastmod>
      <changefreq>monthly</changefreq>
      <priority>0.8</priority>
   </url>
</urlset>
```

- Limited number of links (50k) and volume (50MB) of sitemap files
  - Can be extended using a sitemaps index file - TOC of sitemaps files
- Generally located in web site "root" directory; uses robots.txt

Schema.org Metadata

- Structured metadata about data on the Internet

Schema.org Metadata

- Structured metadata about data on the Internet
- Dataset is a subset of schema.org metadata - "A body of structured information describing some topic(s) of interest."
  - Google dataset search tool now available with schema.org metadata

Schema.org Metadata

- Structured metadata about data on the Internet
- Dataset is a subset of schema.org metadata - "A body of structured information describing some topic(s) of interest."
  - Google dataset search tool now available with schema.org metadata
- Simple JSON-LD (Linked Data) protocol

```
{
  "@context": "https://json-ld.org/contexts/person.jsonld",
  "@id": "http://dbpedia.org/resource/John_Lennon",
  "name": "John Lennon",
  "born": "1940-10-09",
  "spouse": "http://dbpedia.org/resource/Cynthia_Lennon"
}
```

Schema.org Metadata

- Structured metadata about data on the Internet
- Dataset is a subset of schema.org metadata - "A body of structured information describing some topic(s) of interest."
  - Google dataset search tool now available with schema.org metadata
- Simple JSON-LD (Linked Data) protocol*

```
{
  "@context": "https://json-ld.org/contexts/person.jsonld",
  "@id": "http://dbpedia.org/resource/John_Lennon",
  "name": "John Lennon",
  "born": "1940-10-09",
  "spouse": "http://dbpedia.org/resource/Cynthia_Lennon"
}
```

*Nothing is ever simple; see https://json-ld.org/ for more information about JSON-LD

How does EDI use sitemaps.org and schema.org metadata?

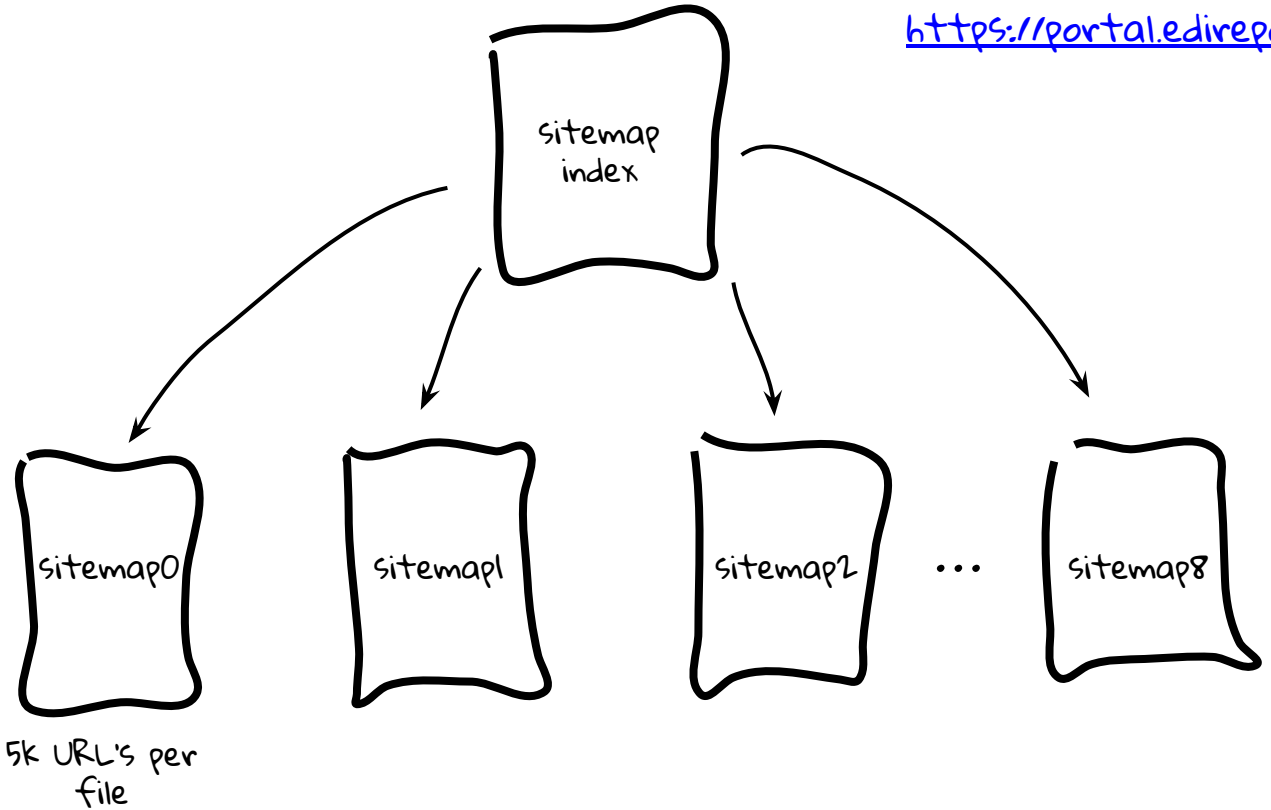
1. EDI generates sitemaps.org metadata for all "latest revision" data package landing pages
   a. published to https://portal.edirepository.org
   b. refreshed hourly

How does EDI use sitemaps.org and schema.org metadata?

1. EDI generates sitemaps.org metadata for all "latest revision" data package landing pages
   a. published to https://portal.edirepository.org
   b. refreshed hourly
2. EDI generates schema.org metadata for all data package landing pages on demand
   a. embeds in <head> element of all landing pages as script type json-ld
   b. accessible at https://seo.edirepository.org

EDI AND SITEMAPS.ORG

https://portal.edirepository.org/



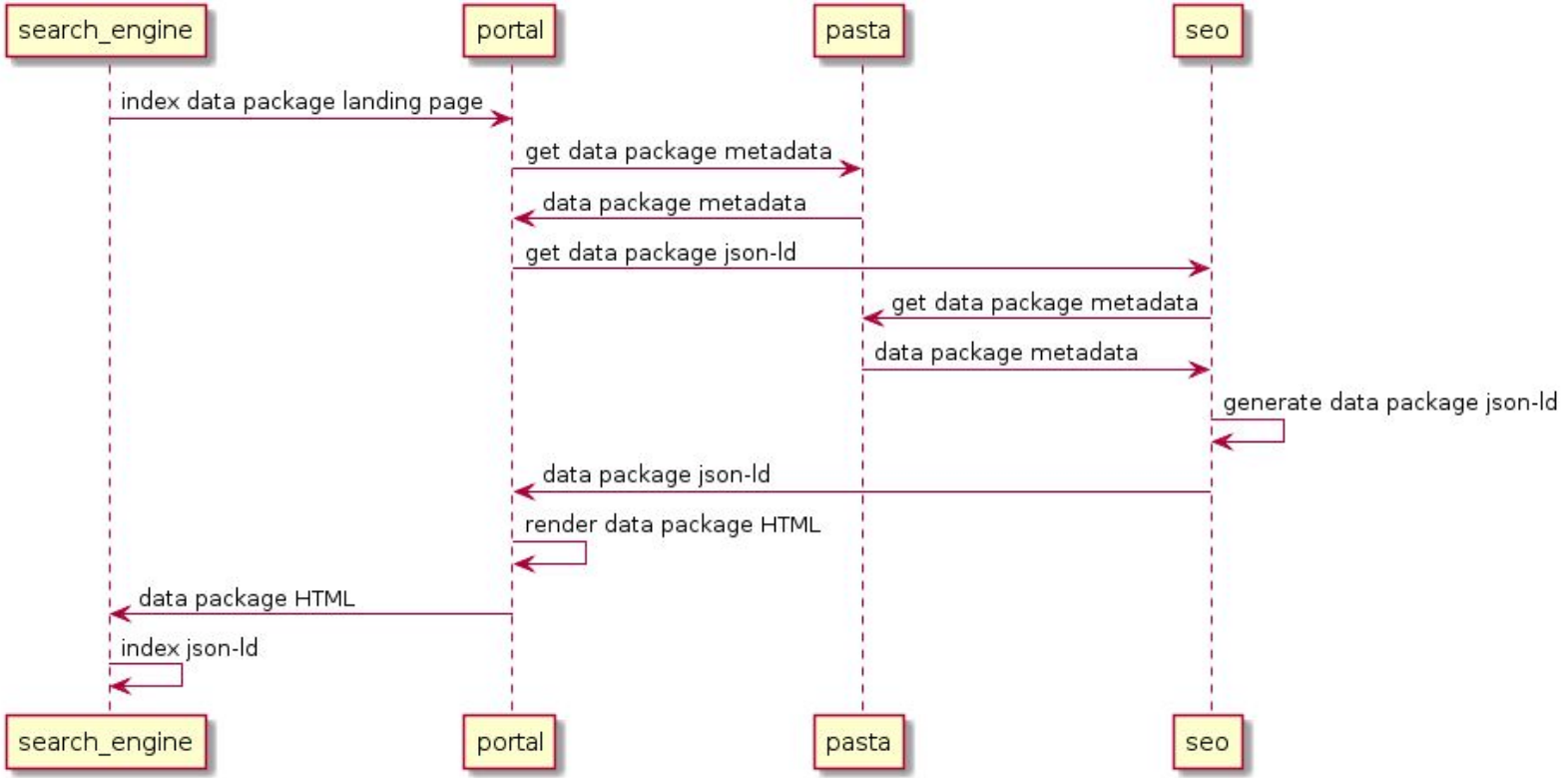
SITEMAP_INDEX.XML

```
<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <sitemap>
    <loc>https://portal.edirepository.org/sitemap0.xml</loc>
  </sitemap>
  <sitemap>
    <loc>https://portal.edirepository.org/sitemap1.xml</loc>
  </sitemap>
  <sitemap>
    <loc>https://portal.edirepository.org/sitemap2.xml</loc>
  </sitemap>
  .
  .
  .
  <sitemap>
    <loc>https://portal.edirepository.org/sitemap8.xml</loc>
  </sitemap>
</sitemapindex>
```

sitemap0.xml

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>
      https://portal.edirepository.org/nis/mapbrowse?scope=ecotrends&identifier=1&revision=2
    </loc>
  </url>
  .
  .
  .
  <url>
    <loc>
      https://portal.edirepository.org/nis/mapbrowse?scope=ecotrends&identifier=14498&revision=2
    </loc>
  </url>
</urlset>
```

EDI AND SCHEMA.ORG



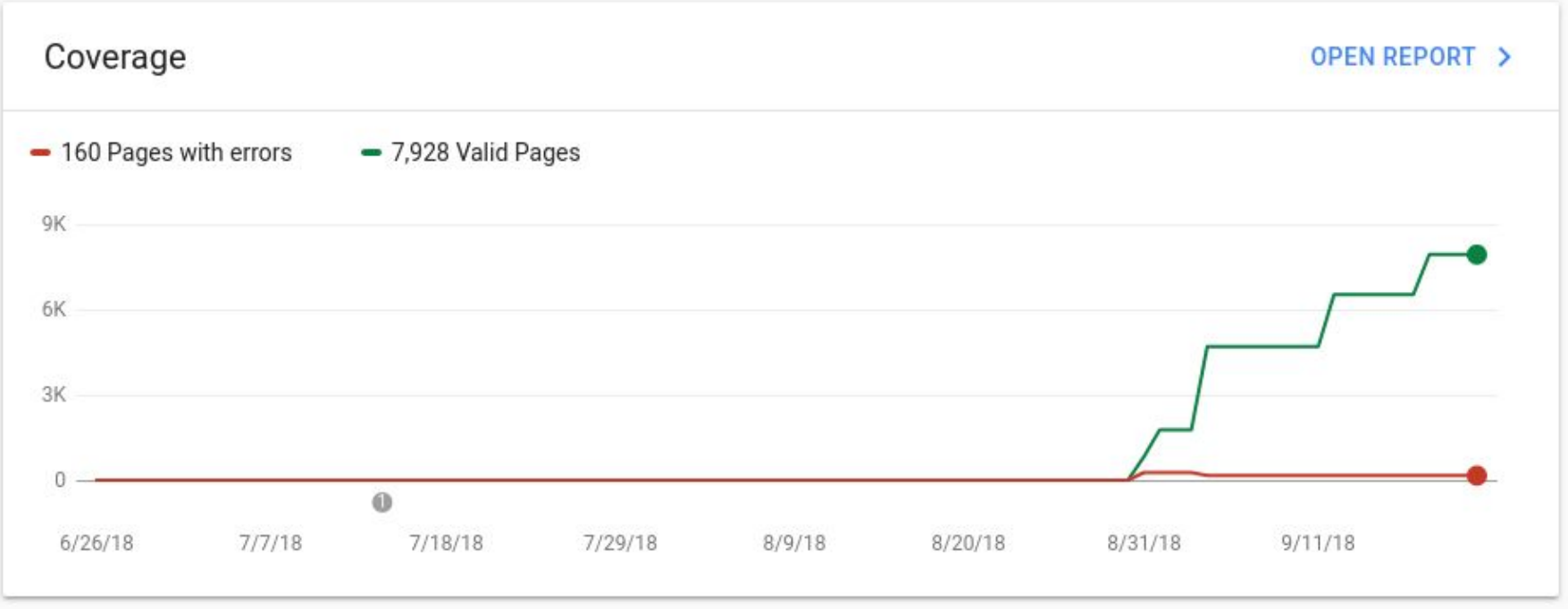
https://seo.edirepository.org/seo/schema/dataset?pid=ecotrends.1.2

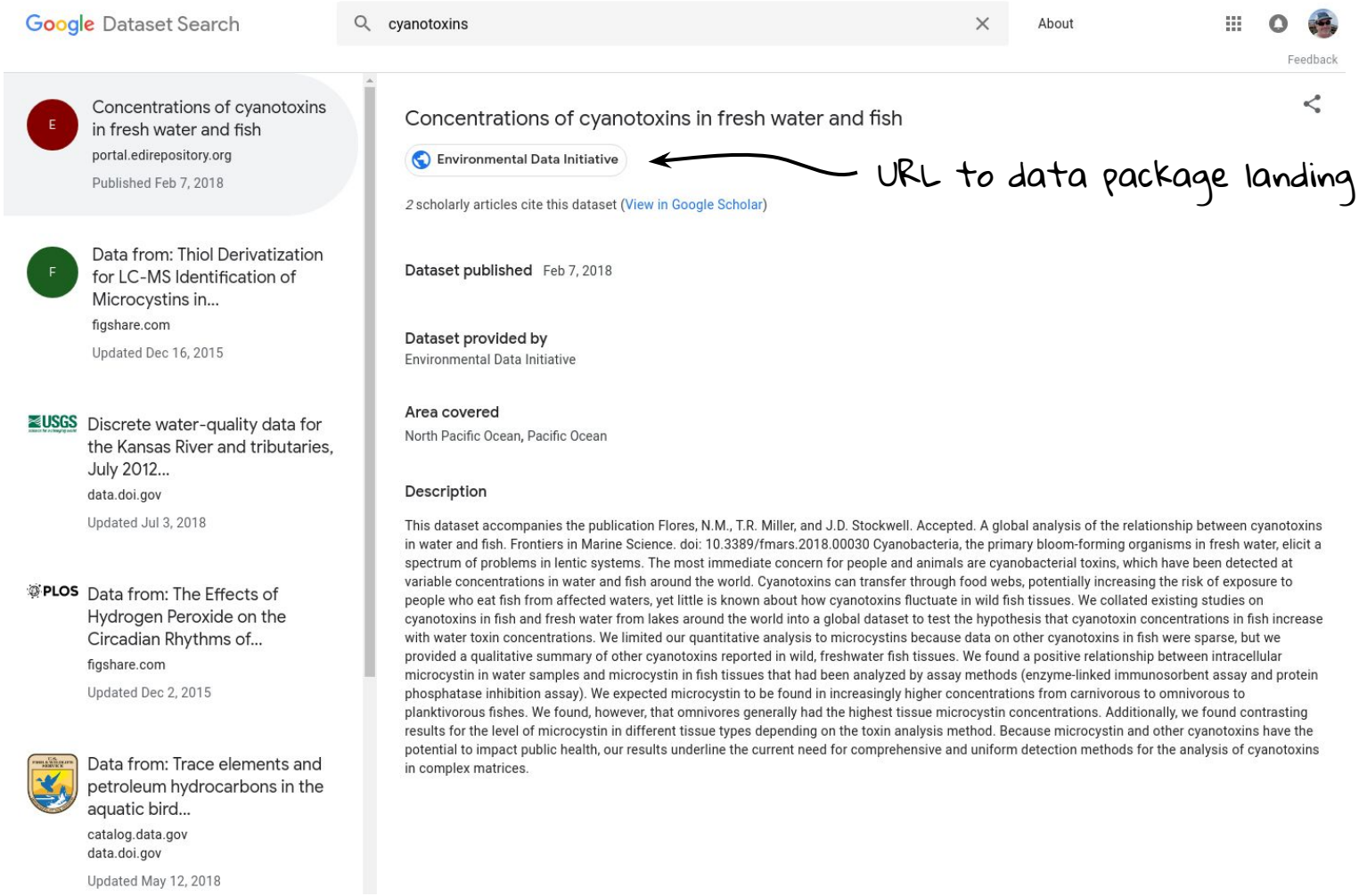

```
{
  "@context": "http://schema.org",
  "@type": "Dataset",
  "name": "H. J. Andrews Experimental Forest site, station Clearcut section of Mack Creek, study of animal
  ...
  "url": "https://portal.edirepository.org/nis/metadataviewer?packageid=ecotrends.1.2",
  "publisher": {
    "@type": "Organization",
    "@id": "https://environmentaldatainitiative.org",
    "name": "Environmental Data Initiative",
    ...
    "url": "https://environmentaldatainitiative.org",
    "email": "info@environmentaldatainitiative.org"
  },
  "description": "The EcoTrends project was established...,
  "datePublished": "2014-10-24",
  "identifier": "doi:10.6073/pasta/930011a435765663182697ad04d147c9"
...
}
```

https://portal.edirepository.org/nis/mapbrowse?scope=ecotrends&identifier=1&revision=2

```
<!DOCTYPE html>
<html lang="en">
<head>
...
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Dataset",
  "name": "H. J. Andrews Experimental Forest site, station Clearcut section of Mack Creek, study of animal
  ...
  "url": "https://portal.edirepository.org/nis/metadataviewer?packageid=ecotrends.1.2",
  "publisher": {
    "@type": "Organization",
    "@id": "https://environmentaldatainitiative.org",
    "name": "Environmental Data Initiative",
    ...
    "url": "https://environmentaldatainitiative.org",
    "email": "info@environmentaldatainitiative.org"
  },
  "description": "The EcoTrends project was established...,
  "datePublished": "2014-10-24",
  "identifier": "doi:10.6073/pasta/930011a435765663182697ad04d147c9"
...
}
</script>
</head>
<body>
```

How does SEO affect data stored in the EDI Data Repository?




Google Dataset Search
cyanotoxins
About
Feedback
Concentrations of cyanotoxins in fresh water and fish
portal.edirepository.org
Published Feb 7, 2018
Data from: Thiol Derivatization for LC-MS Identification of Microcystins in...
figshare.com
Updated Dec 16, 2015
Discrete water-quality data for the Kansas River and tributaries, July 2012...
data.doi.gov
Updated Jul 3, 2018
Data from: The Effects of Hydrogen Peroxide on the Circadian Rhythms of...
figshare.com
Updated Dec 2, 2015
Data from: Trace elements and petroleum hydrocarbons in the aquatic bird...
catalog.data.gov
data.doi.gov
Updated May 12, 2018
Concentrations of cyanotoxins in fresh water and fish
Environmental Data Initiative
URL to data package landing page
2 scholarly articles cite this dataset (View in Google Scholar)
Dataset published Feb 7, 2018
Dataset provided by
Environmental Data Initiative
Area covered
North Pacific Ocean, Pacific Ocean
Description
This dataset accompanies the publication Flores, N.M., T.R. Miller, and J.D. Stockwell. Accepted. A global analysis of the relationship between cyanotoxins in water and fish. Frontiers in Marine Science. doi: 10.3389/fmars.2018.00030 Cyanobacteria, the primary bloom-forming organisms in fresh water, elicit a spectrum of problems in lentic systems. The most immediate concern for people and animals are cyanobacterial toxins, which have been detected at variable concentrations in water and fish around the world. Cyanotoxins can transfer through food webs, potentially increasing the risk of exposure to people who eat fish from affected waters, yet little is known about how cyanotoxins fluctuate in wild fish tissues. We collated existing studies on cyanotoxins in fish and fresh water from lakes around the world into a global dataset to test the hypothesis that cyanotoxin concentrations in fish increase with water toxin concentrations. We limited our quantitative analysis to microcystins because data on other cyanotoxins in fish were sparse, but we provided a qualitative summary of other cyanotoxins reported in wild, freshwater fish tissues. We found a positive relationship between intracellular microcystin in water samples and microcystin in fish tissues that had been analyzed by assay methods (enzyme-linked immunosorbent assay and protein phosphatase inhibition assay). We expected microcystin to be found in increasingly higher concentrations from carnivorous to omnivorous to planktivorous fishes. We found, however, that omnivores generally had the highest tissue microcystin concentrations. Additionally, we found contrasting results for the level of microcystin in different tissue types depending on the toxin analysis method. Because microcystin and other cyanotoxins have the potential to impact public health, our results underline the current need for comprehensive and uniform detection methods for the analysis of cyanotoxins in complex matrices.

Other initiatives using sitemaps.org and schema.org

- EarthCube Project 418
  - Extend the CDF Registry Working Group methodology to data holdings
  - Investigate usage of schema.org/Dataset for Data Facilities
  - Implement schema.org/Dataset with geoscience specific vocabularies
  - Assist Data Facilities with adoption of metadata implementation
  - Develop a cloud-based software stack for crawling, indexing, and access
  - Develop sample user interfaces for accessing cloud-based software stack
- DataONE
  - Publishing sitemaps.org and schema.org metadata on https://search.dataone.org
  - Investigating use of sitemaps.org and schema.org metadata for Member Node harvesting
  - Develop slender node approach for light-weight Member Node registration

Take Aways...

- Evolving technology
- Low-barrier to support for data repositories
- Seamless for data providers
- Search engines have much greater resources for perfecting search
- Stay tuned for future improvements

Thank you

Powered By

PASTA