



Schema.org Metadata: NCEI Implementation Status

John Relph
NOAA National Centers for Environmental Information

July 17, 2018
ESIP Summer Meeting 2018



History of Schema.org/Dataset at NOAA

July 2014, ESIP Summer Meeting, Schema.org Hackathon

- I developed initial implementation of Schema.org JSON-LD markup and deployed at NODC

November 2017, Google meets with NOAA

- Google reviewed NCEI-MD implementation and suggested a few improvements.

January 2018, ESIP Winter Meeting

- Google plenary and second NOAA meeting with Google

March 2018, NOS/NMFS/NCEI fully deploy Schema.org/Sitemap.xml

- InPort now provides full Schema.org metadata for NOS/NMFS datasets; NCEI implementation handles all NCEI datasets



Standards and Best Practices Matter

- Schema.org/Dataset metadata was straightforward to implement for NODC collection metadata because they were consistent
 - Author names were represented consistently in the ISO
 - Online data access links were represented consistently
 - Used boilerplate for much of the information

Standards and Best Practices Matter

- Schema.org/Dataset metadata are less consistent for NCEI metadata because there is less consistency of practice across the multiple offices
 - Authors for non-NODC citations were derived from list of all contributors having certain “creator” roles
 - Online data access links were inconsistent across the various NCEI offices
 - Organizational contacts were even less consistent



Standards and Best Practices Matter

- In order to improve consistency of Schema.org/Dataset metadata, NCEI needed to make changes to their ISO implementations
 - Authors for citations are now taken from the contributors having role “author”
 - Working on NCEI “ISO Template” to standardize Online Data Access and documentation links, organizational contacts, etc.
- Schema.org/Dataset metadata can only be as good as the ISO metadata!

External Review can help

```
"identifier" : [  
  {  
    "value" : "doi:10.3334/CDIAC/OTG.NDP047",  
    "propertyID" : "Digital Object Identifier (DOI)",  
    "@type" : "PropertyValue"  
  },  
  {  
    "value" : "gov.noaa.nodc:0000071",  
    "propertyID" : "NCEI Dataset Identifier",  
    "@type" : "PropertyValue"  
  }  
],
```

- Google representative reviewed our Schema.org/Dataset implementation
 - Noticed that we were not representing DOI information
 - Suggested using Schema.org/identifier
 - I updated our implementation to provide a list of identifiers if a given dataset had a DOI
 - Also suggested building a sitemap



Internal Review

- I provided my hacky code to NOS, they used it as a reference to implement Schema.org/Dataset metadata for NOS and NMFS datasets
 - This triggered a discussion about ongoing datasets with an indeterminate end date, for example, 2014-09-22/present
 - ISO 8601 has no representation for such an interval
 - There are various ISO 8601 drafts, but none have been ratified
 - Still some work to be done in Schema.org and ISO 8601 to represent an indeterminate interval

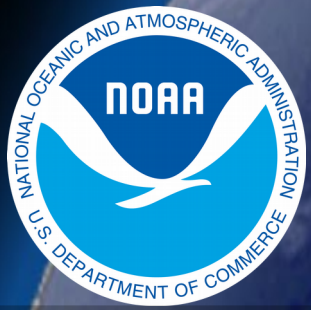
Internal Review

- I created an ISO to Schema.org/Dataset crosswalk
 - More discussion was triggered about identifiers, publication and creation dates, author names for citation, etc.
- As I created the crosswalk and when I later reviewed the crosswalk, I noticed a few things which could be improved
 - Schema.org/Dataset/image could be a URL or an ImageObject, ImageObject includes a description
 - spatial and temporal are now spatialCoverage and temporalCoverage
- Multiple and ongoing reviews can improve implementation



How to Measure Effectiveness

- It was unclear for some time whether Schema.org/Dataset metadata improved search results
- Implementing sitemap and robots.txt, as suggested by Google, seems to have helped
- Anecdotal evidence suggests that more NCEI datasets are now findable in Google, presumably because the sitemap refers crawlers to the complete set of NCEI dataset metadata
- Thus, we recommend using robots.txt and sitemap.xml for best results



Any Questions?

John.Relph@noaa.gov

National Oceanic and Atmospheric Administration | NOAA Satellite and Information Service



National Centers for Environmental Information

Resources

- **NOAA ISO 19115-2 to Schema.org/Dataset crosswalk:**
<https://docs.google.com/spreadsheets/d/10jDiXUB6vD9OfWIB42NrrDelZCG1jpJx0j93mIJVv8/edit?usp=sharing>
- **Schema.org/Dataset:**
<http://schema.org/Dataset>
- **Google: Build, Test, and Release Your Structured Data:**
<https://developers.google.com/search/docs/guides/prototype>
- **Sitemaps XML Format:**
<https://www.sitemaps.org/protocol.html>