Project 418

Using web architecture patterns to address discovery and facilitate approaches to Open and FAIR

Placing data in context

Overview

A general overview of P418

FYI: 418 == April 2018, our project end date

Background

EarthCube CDF Registry Working Group

Focused on elements of facility metadata

Collaboration with RE3Data

Elements included exploring a pattern around self hosted structured data extending schema.org/Organization

Structured data harvested via the web

Motivation

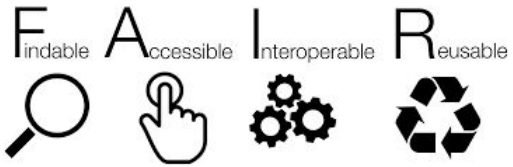
The existence of schema.org/Dataset was known and being used by some RWG members

There was interest in applying experience in RWG to facility datasets

There was also known interest and use of this pattern elsewhere in the community

Drivers & Philosophy

FAIR patterns:



Enabling FAIR Data in the Earth and Space Sciences (Community lead, AGU convened)

How can we help address “F” in a scalable, standards based and easy to implement manner?

Image credit: wikimedia commons

Build a sustainable practice on web standards

Web architecture based

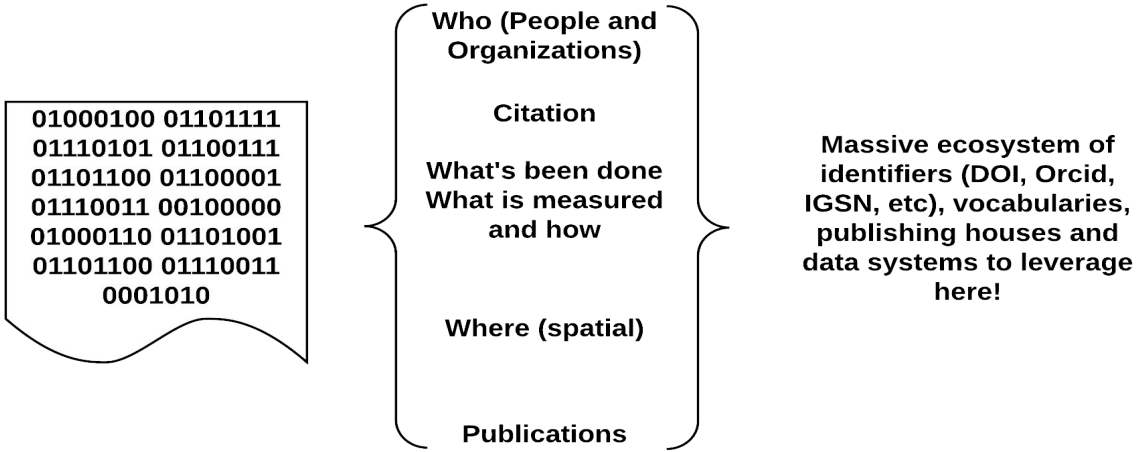
- Leverage web native protocols & patterns that are all around us

Leverage the known and existing

- Web publishing workflows of providers
- Publishing patterns of the web
- Developers tools & libraries
- Access patterns of the consumers

Leveraging

P418 Helps address:



Overall: Helps place data in context

See references for sources (Google, BioShare, EarthCube, COPDESS, Enabling Fair Data)

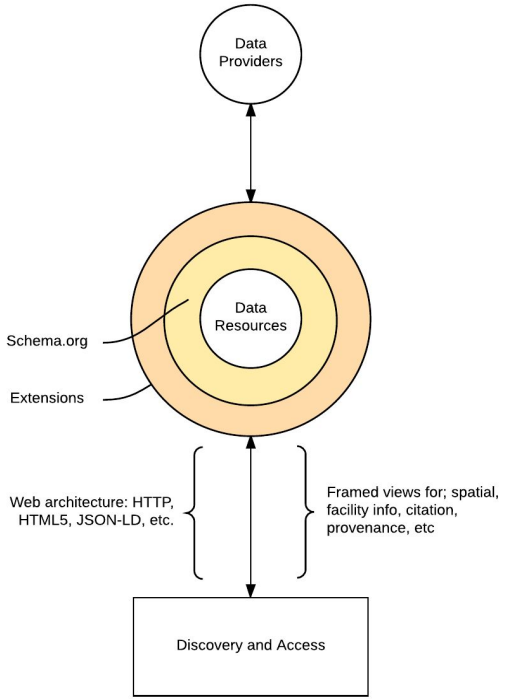
P418: Implementation

SCOPE: Working with a set of NSF data facilities to *demonstrate publishing approaches* for schema.org/Dataset and extensions
(special thanks to them!)

Use schema.org as a base vocabulary with extensions:

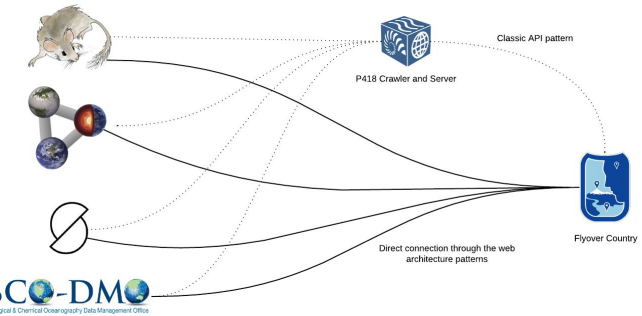
1. Connecting to existing vocs (RE3, GeoLink, OntoSoft, Geoscience Ontology, DCAT, etc)
2. This is a key point: *The patterns provides a means to operationalize this mechanism for data context inside data facilities in a manner aligned with existing web publishing workflows*

SCOPE: *Implement harvesting and interface packages* to further explore the full pipeline and provide feedback.



P418: Principles over Project

- Anyone can take this approach concept and implement it!
  - Example of Flyover country and *disintermediating* P418 implementation packages
- Reduce a priori knowledge needed by all actors (facilities, developers, scientists)
- Leveraging existing work inside and outside EarthCube



P418: Status at T - 3 months

Engaged in the community: Google Research, EarthCube (CDF, RCNs, etc) ESSO PAT team, Enabling FAIR Data, and more

Working with initial data providers (BCO-DMO, LinkedEarth, Neotoma, OpenCore, more coming)

- Vocabulary elements and best practices (reference docs and voc repos)
- Developing connections with existing terms and vocabularies
  - GeoLink, IGSN, PROV, EarthCollab, DCAT, others…
- Exploring impact on web publishing workflows for sites

Implementation elements (used to provide feedback to the approach)

- Harvesting code (All at github, see references)
- User Interfaces GeoDex, Notebooks

Growing Community of Interest

- Google Research
  - data search tool based on this approach in the works
  - ⅓ of web is using structured data in some form
- DataCite
  - APIs for schema.org views
  - RE3data connection
- Bioschemas
- Enabling Fair Data Project
- EarthCube
  - CDF Registry Working Group in connection with RE3Data
  - ESSO Project 418 Data Partners
- Federal interest: NOAA, USGS, NASA
- International: Pangea, Marine.ie, more….
- RDA task force might be in the works???

P418 provider partners

Key/foundational partners in the approach

Also good people to talk to for their POV and opinion.

* Type schema.org/Dataset providers (approx 60K datasets currently)

BCO-DMO*
IEDA
IRIS
Linked Earth*
Martha's Vineyard Coastal Observatory
Neotoma*
Open Core*
Open Topography
R2R
UNAVCO
UNIDATA

Conclusion

Place Data in Context

Enable Value add opportunities

Principles over Project

Use web architecture patterns in a sustainable and scalable manner

- Eric Lingerfelt — EarthCube Technical Officer
- Douglas Fils — Ocean Leadership "data and stuff"
- Adam Shepherd — BCO-DMO Technical Director

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), supported by National Science Foundation grant number ACI-1548562.

POWERED BY XSEDE Extreme Science and Engineering Discovery Environment

EARTHCUBE TRANSFORMING GEOSCIENCES RESEARCH

UCAR UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

Resources and References

References used in this presentation:

- Enabling FAIR Data Project http://www.copdess.org/home/enabling-fair-data-project
- Schema.org http://schema.org/docs/extension.html
- BioSchema https://github.com/BioSchemas & https://f1000research.com/slides/5-2284
- EarthCube CDF Registry working group
  - https://github.com/fils/CDFRegistryWG
  - https://repograph.net/html/webslides/decks/cdfrwg.html#slide=1
- ESIP Lab Provisium https://github.com/ESIPFed/provisium
- EarthCube Project 418 https://github.com/earthcubearchitecture-project418
- Geoscience Ontology http://geoscienceontology.org/
- OntoSoft http://www.ontosoft.org/
- GeoLink http://www.geolink.org/
- COPDESS http://www.copdess.org/


EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH
UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH
NSF

Technical

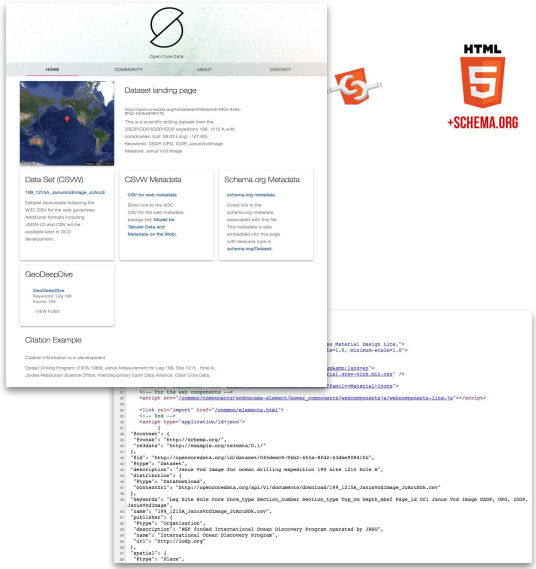
A few more technical slides for when needed

Overview


Publishing
Harvesting
Value Add
Userspace
Sustainable practices
Implementations & Uses

Publishing/Basics

- Structured data in HTML (JSON-LD with type schema.org/Dataset)
- Need to be able to work script header tag into documents
- Recommended to use a sitemap for publishing
- Assessing impact on websites
  - A real need for performance and optimization of sites if crawling becomes an access method
  - Content negotiate for JSON-LD?



Publishing/Basics (HTML + JSON-LD)

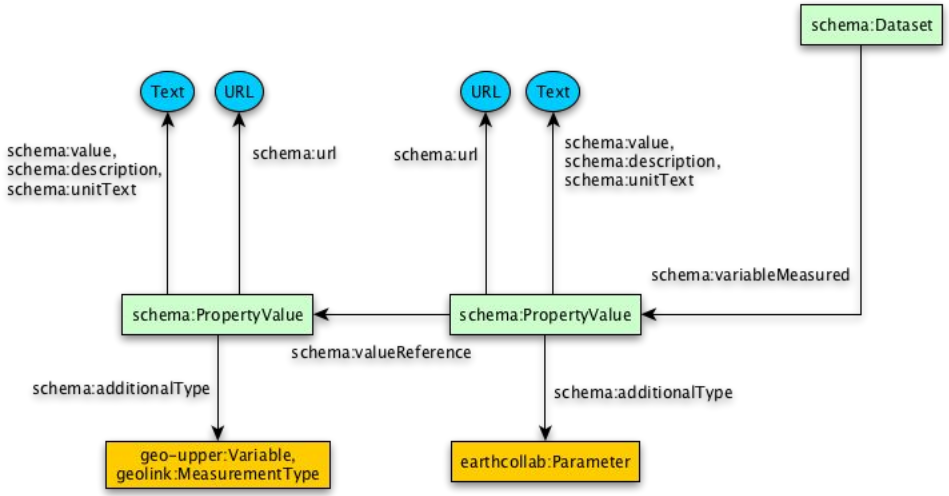
```
<html lang="en">

<head>
    <meta charset="utf-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="description" content="A portfolio template that uses Material Design Lite.">
    <meta name="viewport" content="width=device-width, initial-scale=1.0, minimum-scale=1.0">
    <title>Open Core Data</title>
    <link rel="stylesheet" href="https://fonts.googleapis.com/css?
family=Roboto:regular,bold,italic,thin,light,bolditalic,black,medium&amp;lang=en">
    <link rel="stylesheet" href="https://code.getmdl.io/1.1.3/material.grey-pink.min.css" />
    <link rel="stylesheet" href="/common/styles.css" />
    <link rel="stylesheet" href="https://fonts.googleapis.com/icon?family=Material+Icons">
    <!-- For the web components -->
    <script src="/common/components/ocdparams-element/bower_components/webcomponentsjs/webcomponents-lite.js"></script>

    <link rel="import" href="/common/elements.html">
    <!-- End -->
    <script type="application/ld+json">
          {
 "@context": {
  "@vocab": "http://schema.org/",
  "re3data": "http://example.org/re3data/0.1/"
 },
 "@id": "http://opencoredata.org/id/dataset/045deec9-94b2-445a-8fd2-43dbe90841fb",
 "@type": "Dataset",
 "description": "Janus Vcd Image for ocean drilling expedition 199 site 1215 hole A",
 "distribution": {
  "@type": "DataDownload",
  "contentUrl": "http://opencoredata.org/api/v1/documents/download/199_1215A_JanusVcdImage_JcAruSDk.csv"
 },
 "keywords": "Leg Site Hole Core Core_type Section_number Section_type Top_cm Depth_mbsf Page_id Url Janus Vcd Image DSDP, OPD, IODP,
JanusVcdImage",
 "name": "199_1215A_JanusVcdImage_JcAruSDk.csv",
 "publisher": {
  "@type": "Organization",
  "description": "NSF funded International Ocean Discovery Program operated by JRSO",
  "name": "International Ocean Discovery Program",
  "url": "http://iodp.org"
 },
 "spatial": {
  "@type": "Place",
```

Publishing/Vocabularies

https://github.com/earthcubearchitecture-project418/p418Vocabulary

Using schema.org as a basis with a focus on type Dataset. Then providing example and reference implementation of using external vocabularies to address domain specific needs.

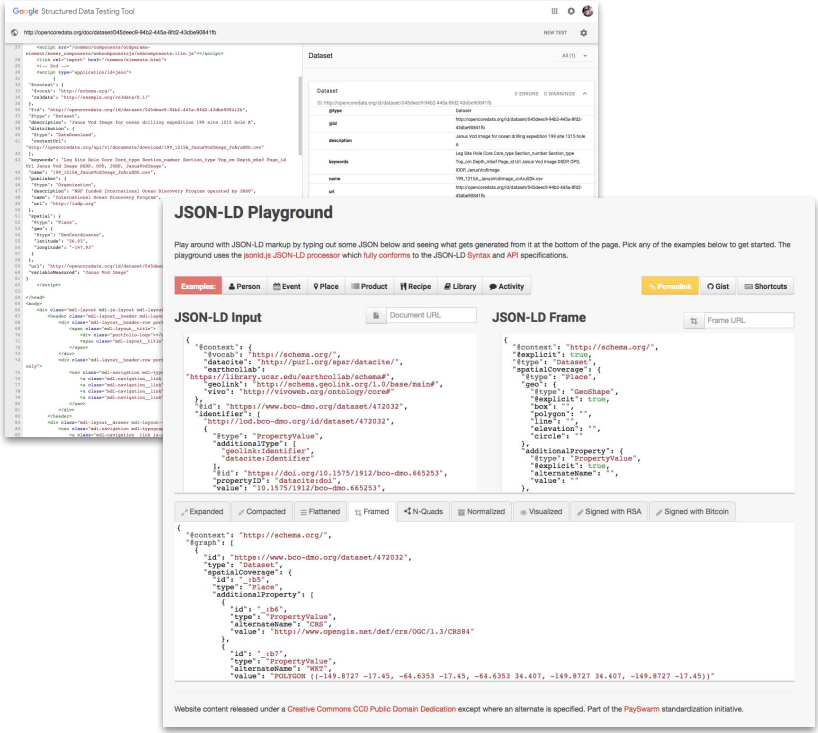
1. To produce quality schema.org markup with additional extensions to schema.org classes to help improve harvesting technologies.
2. Produced markup will pass the Google Structured Data Testing Tool with 0 errors.



Publishing/Tools

Tools and guides:

- P418 guide https://github.com/earthcubearchitecture-project418/p418Docs/blob/master/publishing.md
- Google: Structured data testing tool https://search.google.com/structured-data/testing-tool/u/0/#
- JSON-LD playground https://json-ld.org/playground

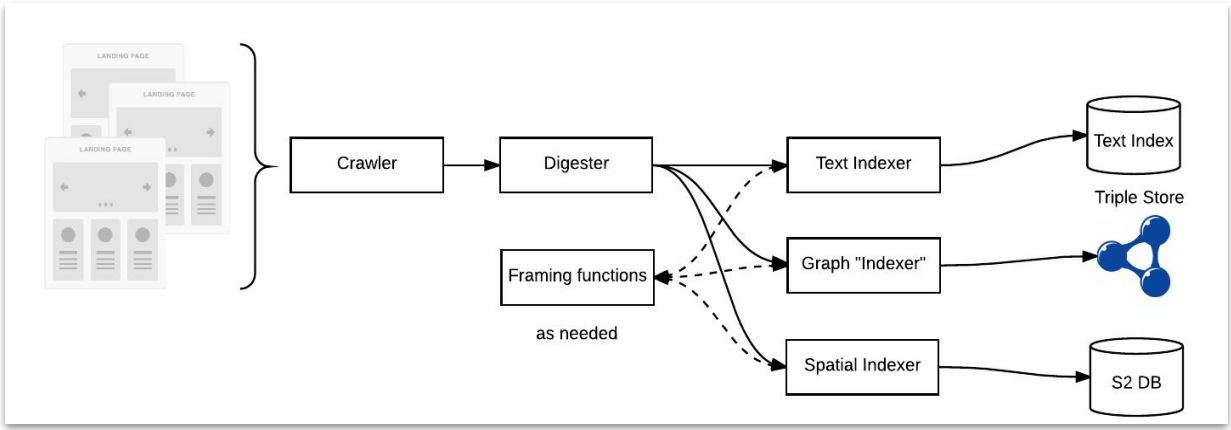


Implementation: Harvesting approach

Crawler at: https://github.com/earthcubearchitecture-project418/crawler

- spatial (geohash)
- index (bleve)
- RDF (blazegraph) (Need to address blank nodes from JSON-LD)
- Temporal (later)

Leveraging JSON-LD framing

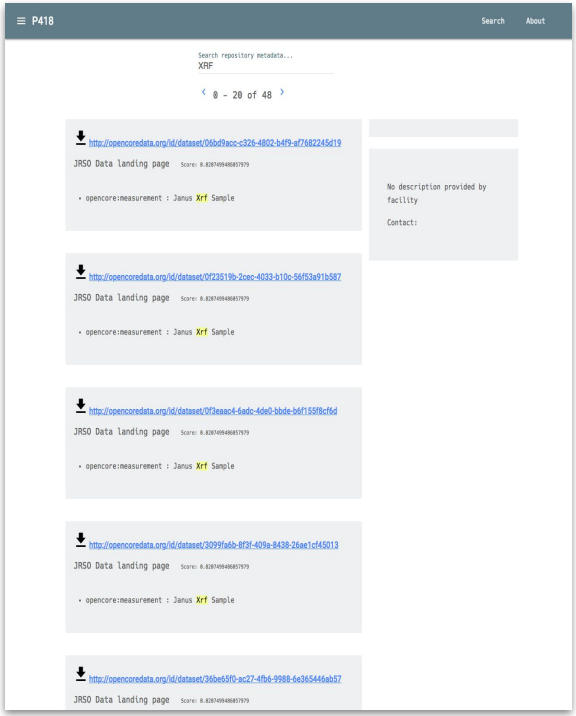


Implementation: Interfaces

The generated indexes are exposed by a collection of APIs

Web implementation of APIs for testing
https://geodex.org

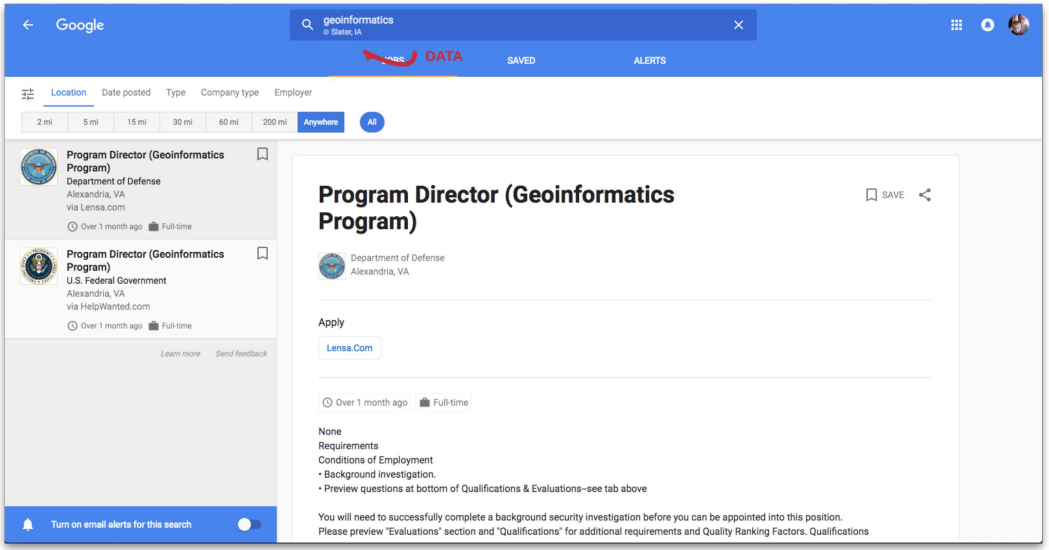
Notebooks
https://github.com/earthcubearchitecture-project418/p418Notebooks



Google Research Interest

Google Data search is likely to be similar to the Google Careers “job” search. Job search is also driven by schema.org structured data.

Research at Google



EarthCube
TRANSFORMING GEOSCIENCES RESEARCH

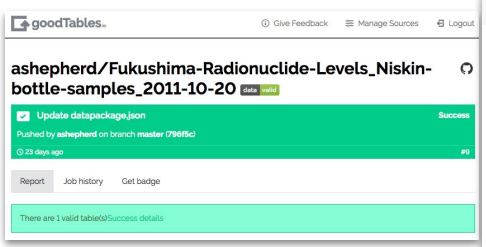
UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

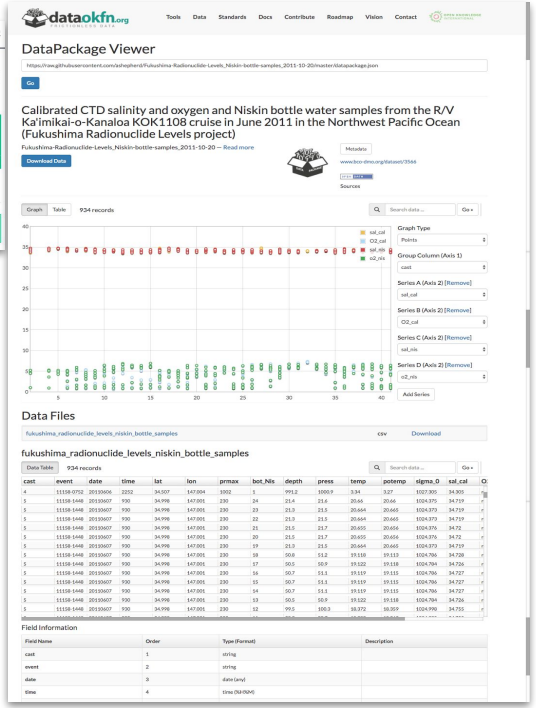
Side Focus: Data packages

FRICTIONLESS DATA
STANDARDS AND TOOLING

https://frictionlessdata.io/

Facilitating connection of data to potential actions and connecting to native data types in R or Python for example





Abstract ID: 236488
Abstract Title: The Frictionless Data Package: Data Containerization for Automated Scientific Workflows
Final Paper Number: IN33C-0139
Presentation Type: Poster
Session Date and Time: Wednesday, 13 December 2017; 13:40 - 18:00
Session Number and Title: IN33C: Facilitating Interdisciplinary Geosciences by Limiting Data Friction: Approaches and Outcomes I Posters
Location: New Orleans Ernest N. Morial Convention Center; Poster Hall D-F

EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

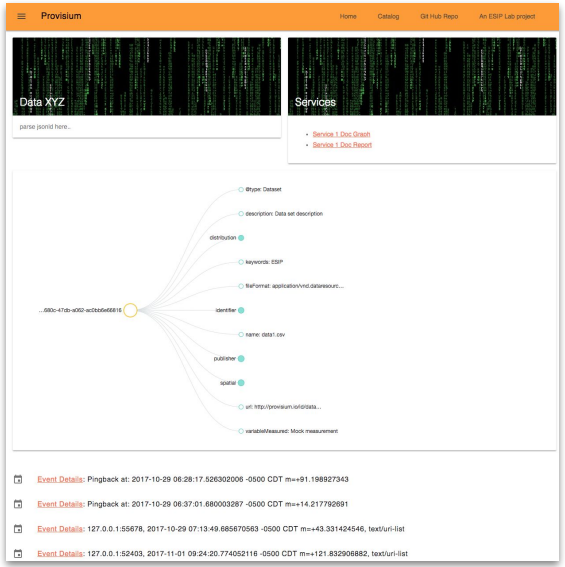
Side Focus: Provenance

ESIP Lab: Provisium

https://github.com/ESIPFed/provisium

- An exploration in implementing the PROV-AQ Note
- Allows for exposing PROV elements via web architecture
- Can be connected in via JSON-LD housing schema.org/Dataset

```
prov:pingback
prov:has_provenance
prov:*
```



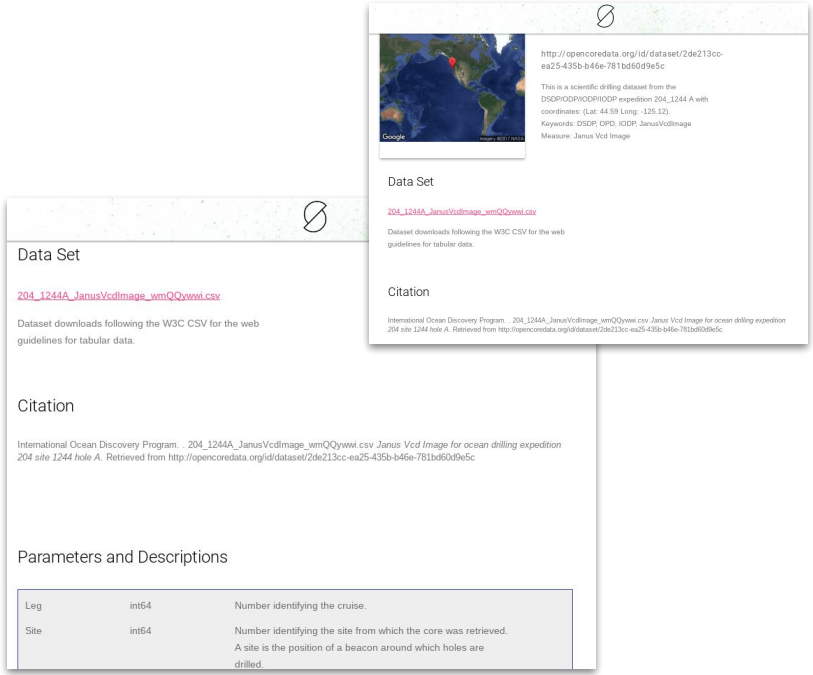
Side Focus: Component Driven Landing Pages

Landing Pages + Web Components

Leveraging machine readable data to generate human focused “snippets”.

Citations, maps, parameter listing and more

- More efficient development
- Shared approaches
- Promoting best practices to a wider set of providers



Conclusion

Place Data in Context

Enable Value add opportunities

Principles over Project

Use web architecture patterns in a sustainable and scalable manner

- Eric Lingerfelt — EarthCube Technical Officer
- Douglas Fils — Ocean Leadership "data and stuff"
- Adam Shepherd — BCO-DMO Technical Director

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), supported by National Science Foundation grant number ACI-1548562.

POWERED BY XSEDE Extreme Science and Engineering Discovery Environment

EARTHCUBE TRANSFORMING GEOSCIENCES RESEARCH

UCAR UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH


EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH
UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH
NSF

Errata slide: Geoinformatics Lattice

RE3
Orchid
Datacite
CodeMeta
etc...

AGU FAIR
COPDESS
Scholix, Hypothesis
etc...

Leveraging the web to build a “data lattice”. A strong, sustainable and persistent interconnection of elements that supports a network of actors and actions on that data. Scoping provenance, citation, annotation, linking and more.