



# 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**

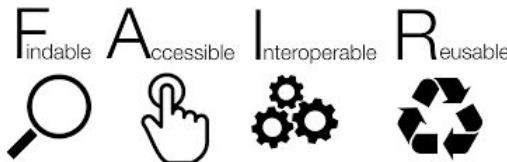
The existence of schema.org/Dataset was known and being used by some RWG members

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)*

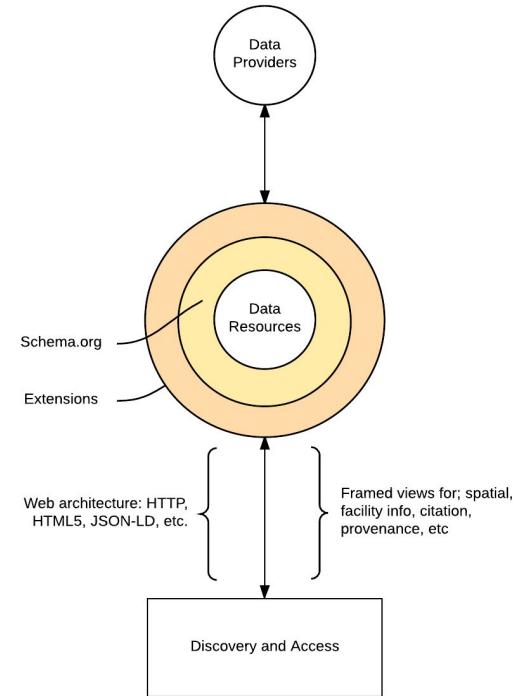
# P418: Implementation

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:

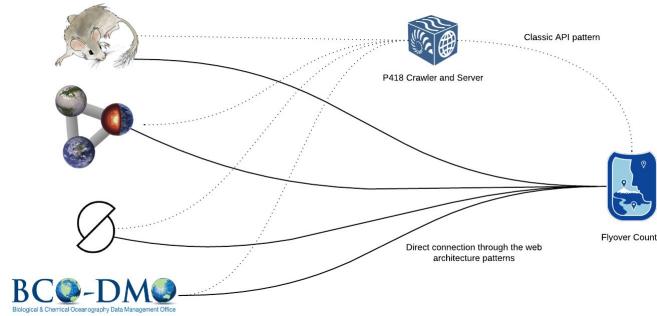
1. Connecting to existing vocs (RE3, GeoLink, OntoSoft, Geoscience Ontology, DCAT, etc)
2. This is a key point: *The patterns provides a means to operationalize this mechanism for data context inside data facilities in a manner aligned with existing web publishing workflows*

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
  - $\frac{1}{3}$  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.*



# Resources and References

## References used in this presentation:

- Enabling FAIR Data Project <http://www.copdесс.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/provisum>
- 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.copdesson.org/](http://www.copdесс.org/)



**EARTHCUBE**  
TRANSFORMING GEOSCIENCES RESEARCH

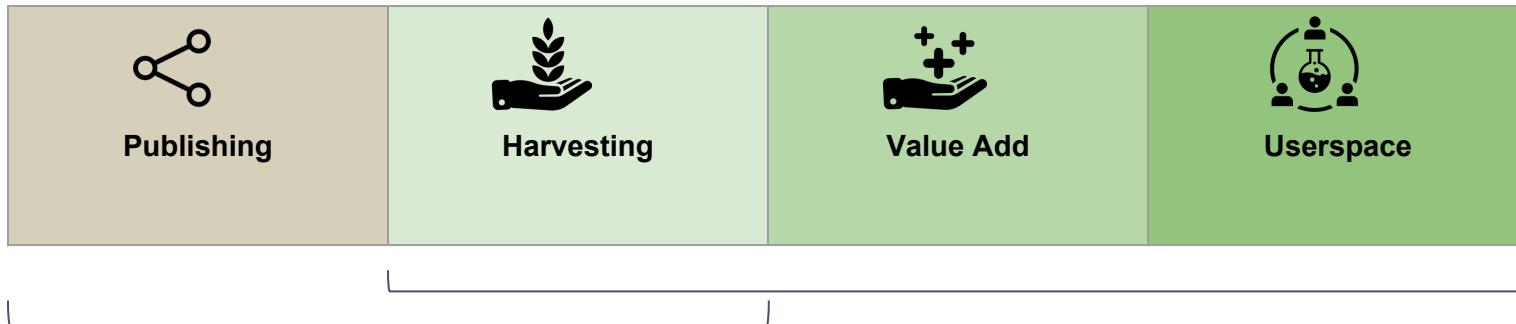
 **UCAR**  
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

 **NSF**

# Technical

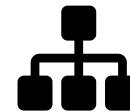
A few more technical slides for when needed

# Overview



# 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?



The screenshot shows a dataset landing page for the Deep Sea Drilling Program (DSDP) Hole 109. The page features a map of the world with a red dot marking the location of the measurement. Below the map, there is a brief description of the measurement: "Citation information is as follows: DSDP Hole 109, Site 109, Hole A; Joides Resolution Science Office, interdisciplinary Earth Sciences, Open Core Data." The page is divided into several sections:

- Data Set (CSVW)**: Includes links for "CSV for web metadata" and "CSV for the web metadata".
- CSVW Metadata**: Includes links for "CSVW for web metadata" and "CSV for the web metadata".
- Schema.org Metadata**: Includes links for "Schema.org metadata" and "Schema.org metadata".
- GeoDeepDive**: Includes links for "GeoDeepDive" and "View Pub".

A "Citation Example" section provides an example of how the data might be cited in a paper, using BibTeX format. The example includes fields like `@dataset`, `title`, `url`, `version`, `format`, and `description`.

# Publishing/Basics (HTML + JSON-LD)

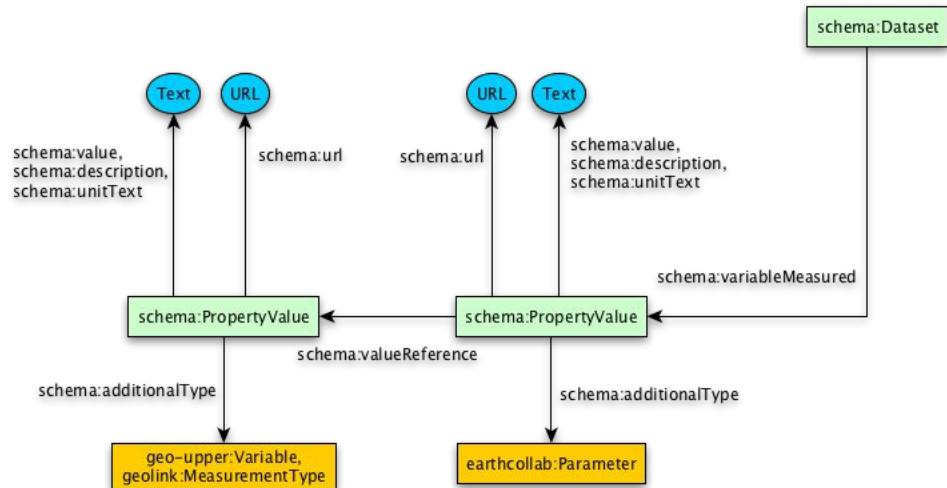
```
19 <html lang="en">
20
21 <head>
22   <meta charset="utf-8">
23   <meta http-equiv="X-UA-Compatible" content="IE=edge">
24   <meta name="description" content="A portfolio template that uses Material Design Lite.">
25   <meta name="viewport" content="width=device-width, initial-scale=1.0, minimum-scale=1.0">
26   <title>Open Core Data</title>
27   <link rel="stylesheet" href="https://fonts.googleapis.com/css?family=Roboto:regular,bold,italic,thin,light,bolditalic,black,medium&lang=en">
28   <link rel="stylesheet" href="https://code.getmdl.io/1.1.3/material.grey-pink.min.css" />
29   <link rel="stylesheet" href="/common/styles.css" />
30   <link rel="stylesheet" href="https://fonts.googleapis.com/icon?family=Material+Icons">
31   <!-- For the web components -->
32   <script src="/common/components/ocdparams-element/bower_components/webcomponentsjs/webcomponents-lite.js"></script>
33
34   <link rel="import" href="/common/elements.html">
35   <!-- End -->
36   <script type="application/ld+json">
37     {
38       "@context": {
39         "@vocab": "http://schema.org/",
40         "re3data": "http://example.org/re3data/0.1/"
41       },
42       "@id": "http://opencoredata.org/id/dataset/045deec9-94b2-445a-8fd2-43dbe90841fb",
43       "@type": "Dataset",
44       "description": "Janus Vcd Image for ocean drilling expedition 199 site 1215 hole A",
45       "distribution": {
46         "@type": "DataDownload",
47         "contentUrl": "http://opencoredata.org/api/v1/documents/download/199_1215A_JanusVcdImage_JcAruSDK.csv"
48       },
49       "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",
50       "name": "199_1215A_JanusVcdImage_JcAruSDK.csv",
51       "publisher": {
52         "@type": "Organization",
53         "description": "NSF funded International Ocean Discovery Program operated by JRSO",
54         "name": "International Ocean Discovery Program",
55         "url": "http://iodp.org"
56       },
57       "spatial": {
58         "@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.

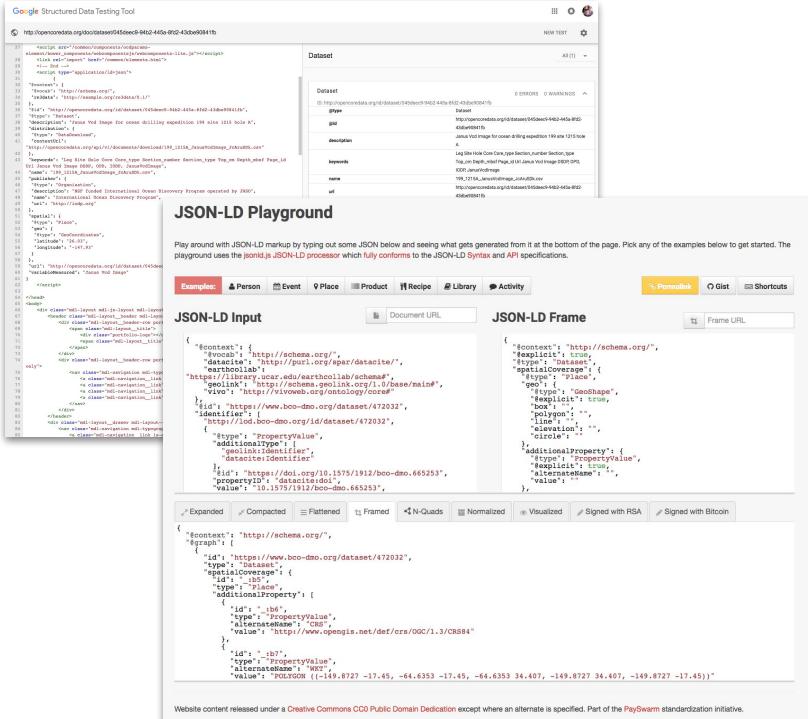
1. To produce quality schema.org markup with additional extensions to schema.org classes to help improve harvesting technologies.
2. Produced markup will pass the [Google Structured Data Testing Tool](#) with 0 errors.



# Publishing/Tools

## Tools and guides:

- P418 guide  
<https://github.com/earthcubearchitect ure-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>



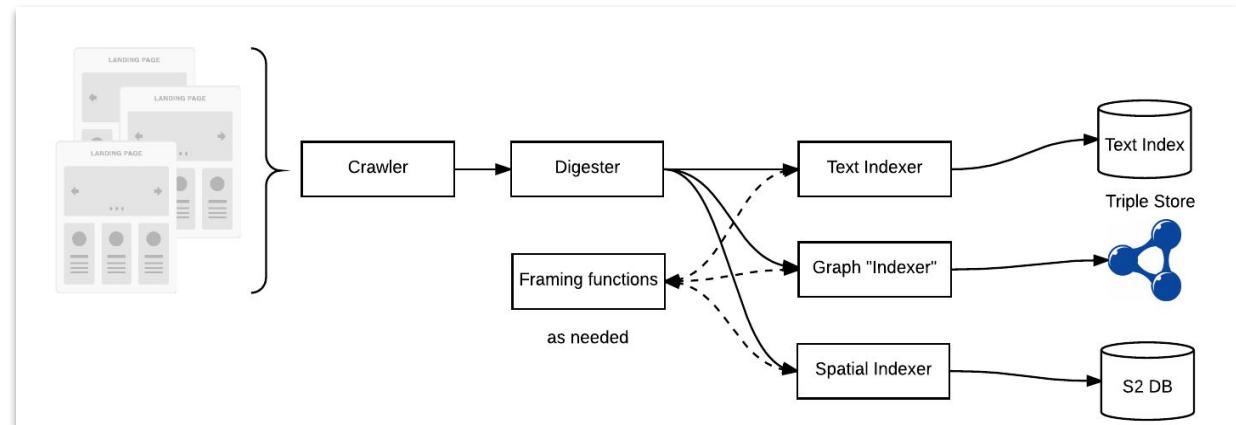
The screenshot shows two panels from the Google Structured Data Testing tool. The left panel displays a JSON-LD document with various schema types like Article, Dataset, and Place. The right panel shows the resulting structured data representation, which includes a 'Dataset' object with properties like 'url', 'name', and 'description'. Below these panels is a JSON-LD playground interface with tabs for Person, Event, Place, Product, Recipe, Library, Activity, and Examples. It also features buttons for 'Permissions', 'Gist', and 'Shortcuts'. A large JSON-LD frame is at the bottom, showing a complex nested structure of objects and properties.

# 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



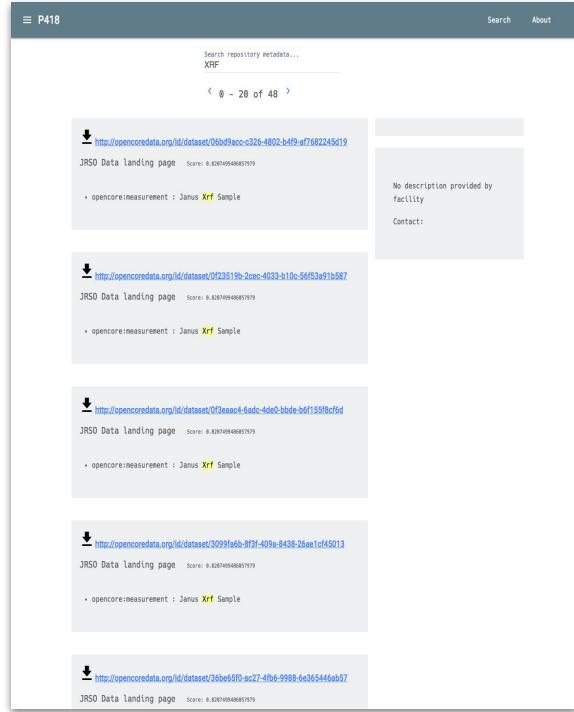
# Implementation: Interfaces

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>



The screenshot shows a web-based interface for managing datasets. At the top, there's a search bar labeled "Search repository metadata..." and a link to "XRF". Below the search bar, it says "0 - 20 of 48". The main area displays five cards, each representing a dataset:

- Card 1:** <http://opencoredata.org/id/dataset/06bd9acc-c324-4802-b4f9-a7f682245d19> (Score: 8.02674948657979)  
+ opencore:measurement : Janus Xrf Sample
- Card 2:** <http://opencoredata.org/id/dataset/0f23519b-2ce0-4033-b10c-56f53a71b587> (Score: 8.02674948657979)  
+ opencore:measurement : Janus Xrf Sample
- Card 3:** <http://opencoredata.org/id/dataset/0f3eaa4-dad0-4de0-bbde-b6f11598bcfd> (Score: 8.02674948657979)  
+ opencore:measurement : Janus Xrf Sample
- Card 4:** <http://opencoredata.org/id/dataset/2099fb0-87f5-409a-8438-26ae1cf45013> (Score: 8.02674948657979)  
+ opencore:measurement : Janus Xrf Sample
- Card 5:** <http://opencoredata.org/id/dataset/260e65f0-ac27-4fb6-9988-6e365446ab57> (Score: 8.02674948657979)  
+ opencore:measurement : Janus Xrf Sample

On the right side of the interface, there are two sections: "No description provided by Facility" and "Contact:".

# 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.



A screenshot of a Google search results page for the query "geoinformatics" with the location set to "St. Paul, IA". The results are filtered under the "DATA" tab. The first result is a job listing for a "Program Director (Geoinformatics Program)" at the Department of Defense in Alexandria, VA, posted via Lensa.com. The second result is for a similar position at the U.S. Federal Government. Both listings show they were posted over 1 month ago and are full-time. The right side of the screen shows a detailed view of the first job listing, including the title, employer, location, and a "Save" button.

# Side Focus: Data packages



<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

The screenshot shows a green success message: "Success" with a "data.json" link. It includes a timestamp ("Pushed by ashepherd on branch master (796fc) ① 23 days ago"), a "Report" button, and a "Get badge" button. Below the message, it says "There are 1 valid table(s) Success details".

The screenshot shows the DataPackage Viewer interface for the dataset "Calibrated CTD salinity and oxygen and Niskin bottle water samples from the R/V Kaiimikai-o-Kanaloa KOK1108 cruise in June 2011 in the Northwest Pacific Ocean (Fukushima Radionuclide Levels project)". It includes a "Download Data" button, a "Metadata" section with links to the dataset's GitHub page and the Data Package JSON file, and a detailed view of the data structure with tables for "cast", "event", "date", "time", "lat", "lon", "pmxx", "bot\_Nls", "depth", "press", "temp", "pbottom", "sigma\_0", and "sal.cal".

# 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:\*

The screenshot shows the Provisium web application. At the top right is the ESIP logo. The main header says "Provisium". Below the header, there are two sections: "Data XYZ" and "Services". The "Data XYZ" section contains a "parse jsonld here..." input field and a "Services" dropdown menu with options "Service 1 Doc Draft" and "Service 1 Doc Report". The central part of the page displays a JSON-LD graph for a dataset. A yellow circle highlights the identifier property, which points to a value. Other properties shown include type, distribution, keywords, mediaType, identifier, name, publisher, spatial, and variableMeasured. At the bottom of the page, there is a list of event details:

- Event Details: Pingback at: 2017-10-29 06:28:17.526302006 -0500 CDT m+=#1.198927343
- Event Details: Pingback at: 2017-10-29 06:37:01.680003287 -0500 CDT m+=#14.217792691
- Event Details: 127.0.0.1:55678, 2017-10-29 07:19:49.685670583 -0500 CDT m+=#43.331424546, text/url-list
- Event Details: 127.0.0.1:52403, 2017-11-01 09:24:20.774052116 -0500 CDT m+=#121.832906882, text/url-list

# 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

The screenshot shows a landing page for a dataset from the International Ocean Discovery Program (IODP). At the top, there is a map of the world with a red dot indicating the location of the expedition. Below the map, the URL is <http://opendatacore.org/id/dataset/2de213cc-ea25-435b-b46e-781bd6d9e5c>. The main content area is divided into sections: "Data Set", "Citation", and "Parameters and Descriptions".  
**Data Set:** A link to [204\\_1244A\\_JanusVcdImage\\_wmQywwi.csv](204_1244A_JanusVcdImage_wmQywwi.csv). A note says: "Dataset downloads following the W3C CSV for the web guidelines for tabular data."  
**Citation:** A note: "International Ocean Discovery Program . 204\_1244A\_JanusVcdImage\_wmQywwi csv Janus Vcd Image for ocean drilling expedition 204 site 1244 hole A. Retrieved from http://opendatacore.org/id/dataset/2de213cc-ea25-435b-b46e-781bd6d9e5c"  
**Parameters and Descriptions:** A table showing parameters:

Leg	int64	Number identifying the cruise.
Site	int64	Number identifying the site from which the core was retrieved. A site is the position of a beacon around which holes are drilled.

# 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.*





**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.