



# Data Discovery and Integration

## An exploration of concepts for IODP

Douglas Fils  
[dfils@oceanleadership.org](mailto:dfils@oceanleadership.org)

# On efforts at [data.oceandrilling.org](http://data.oceandrilling.org)

Fine print: This is a test and NOT (yet) a production service.

How to share: data files, sample level data, images, publications, age models, bug pictures and just about everything we collect... in a way that is compelling and useful.

5th column: Bob, Chris, Cyndy, Stace, Andy, Carla

# Why is this important?

Our community must search multiple locations (inside and outside our program), often with different "keys & concepts". Then align the differences in data before they even start.

Our developers and data managers have a situation where data and logic is hard to intercompare and maintain and spread between data model space, code, service patterns and frameworks.

You can find it... if you already know its there..... and where to look...

# How we approached this / What we did

## On Federation:

Linked Open Data (LOD) & vocabularies as a basis to explore approaches

RDF + SPARQL

5 star data

Query vs API

## On Discovery:

Differentiate "discovery" from "search"

Encourage 3rd parties to do the discovery and inspection of data

Enable better discovery via Google et al

Work with other related providers to link in URI and vocabulary space

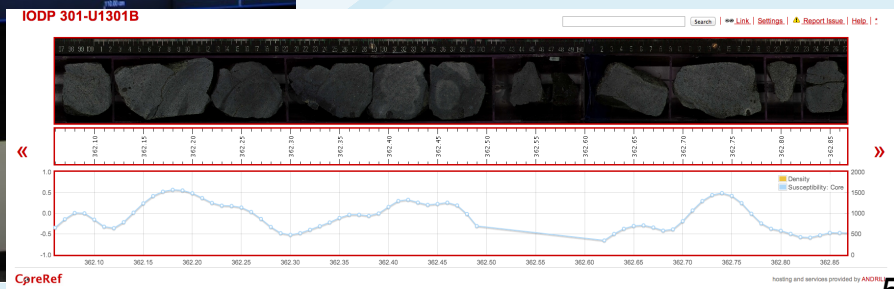
# Federation (provides discovery and exploration)

The image illustrates a federated data environment for ocean drilling research. It features a physical display of sediment core images with a researcher pointing at them, and several browser windows overlaid on top. The windows show data from Scientific Ocean Drilling (IODP) and USIO Lithology Discovery. One window displays a metadata table for 'Section Top: 244.4' with columns for Key and Values. Another window shows a table with columns 'Sand', 'Gravel', and 'Mud'. A blue arrow points from the text '3rd party tools' to the browser windows.

Key	Values
http://xmins.org/foaf/1/name	LDEO
http://purl.org/dc/terms/1/date	11/9/2009 13:6
http://xmins.org/foaf/1/homepage	http://www.ldeo.columbia.edu
http://purl.org/dc/terms/1/source	http://brg.ldeo.columbia.edu/magestrips/0724C0724C27X_Strip0724C27X_2_co_tr.tif
http://purl.org/dc/terms/1/source	http://brg.ldeo.columbia.edu/magestrips/0724C0724C27X_Strip0724C27X_2_co_tr_sc.tif
http://purl.org/dc/terms/1/source	http://brg.ldeo.columbia.edu/magestrips/0724C0724C27X_Strip0724C27X_2_gy_tr.tif
http://purl.org/dc/terms/1/source	http://brg.ldeo.columbia.edu/magestrips/0724C0724C27X_Strip0724C27X_2_gy_tr.tif
http://data.oceandrilling.org/foaf/1/role	C
http://data.oceandrilling.org/foaf/1/sectortop	244.4
http://data.oceandrilling.org/foaf/1/sectiontop	0724C
http://img.org/foaf/1/stripname	62
http://img.org/foaf/1/stripname	1,548,020,849
http://img.org/foaf/1/stripname	558,44,688,2657
http://img.org/foaf/1/stripname	724C27X.tif
http://img.org/foaf/1/stripname	724C27X.tif
http://img.org/foaf/1/stripname	27X
http://img.org/foaf/1/stripname	1.6
http://img.org/foaf/1/stripname	5149,3381
http://img.org/foaf/1/stripname	724
http://img.org/foaf/1/stripname	2
http://img.org/foaf/1/stripname	0724C27X_2

	Sand	Gravel	Mud
0	0	0	100
100	0	0	0
100	0	0	0
100	0	0	0
100	0	0	0

3rd party tools



# Multiple viewing options

Unity Search

Quaried Sources (Government) Meesozoic Planktonic Foraminiferal Working Group

Search for: taxa:Rugoglobigerina

Do Search

Examples site:039 taxa:poro wpt:039:taxa:ava: site:1200 jna:myr\_section geotime:K

Notes: add in search for age models | Central inventory for Samples (need a harvestor) / Publications

Forum Working Group Results

**Rugoglobigerina** *Rugoglobigerina* bubosa is close to the *Rugoglobigerina* macrocephala group than to any other described Upper Cretaceous Foraminifera, particularly to *Rugoglobigerina* macrocephala ornata Brönnimann. The chambers of *Rugoglobigerina* bubosa do not increase in size as rapidly as in *Rugoglobigerina* macrocephala subsp., the test is more rounded in outline, and it also lacks radial ornamentation. Foliolate spirally corded and finely densely corded forms have been found.

**Rugoglobigerina** *Rugoglobigerina* ornata Brönnimann, but rather constant differences in size and development of the marginal pattern justify separate subspecies. The test of *ornata* is larger than that of *macrocephala*, and in addition shows a more pronounced meridional pattern in the adult. It occupies an intermediate position between *Rugoglobigerina* (*Rugoglobigerina*) *macrocephala* Brönnimann and *Rugoglobigerina* (*Rugoglobigerina*) *nugosa* (Plummer) (*Globigerina* *nugosa*, 1927).

The initial portion of this form seems to be identical with that of typical representatives of *Rugoglobigerina* (*Rugoglobigerina*) *nugosa* (Plummer) (*Globigerina* *nugosa*, 1927) and *Rugoglobigerina* (*Rugoglobigerina*) *nachli* Brönnimann.

**Rugoglobigerina** *Rugoglobigerina* reischli

**Rugoglobigerina** *Rugoglobigerina* reischli

**Rugoglobigerina** *Rugoglobigerina*

Scientific Ocean Drilling

Resources

About Working Group

Timeline

Print List

Developed by the Meesozoic Planktonic Foraminiferal Working Group to help users make informed decisions on taxonomic synonymy and validity. It consists of over 470 species records that include original and amended species descriptions, morphologic descriptor lists, synonymy, lithostratigraphic range information, original and new images of testolyses, and SEM images that illustrate the morphologic variability of the species concept. Single and multiple taxonomy, morphology, and time/pace fields are searchable and can be downloaded in a tab-delimited format.

Citation Information: Meesozoic Planktonic Foraminiferal Working Group (Huber, B.T., Coordinator), 2006. Meesozoic Planktonic Foraminiferal Taxonomic Dictionary. www.socdrilling.org

Related Taxa

Width (mm): 310

Record Number: 310

Current: Cretaceous

Genus: Rugoglobigerina

Author: Belford

Date: 1960

Repository: Commonwealth Palaeontological Collection, Canberra, Australia

Cat Number: 2000

Diameter: 0.34 to 0.39

Length (mm): 0.34 to 0.39

Type Location: From the escarp at Pillanes Hill, in the Murchison River area of Western Australia.

Foli Stage: Serranin

Foli Size: 65.6

LAD Stage: Campanian

Last Mm: 60.0

Character Arrangement: Theoretical

Features: Test robust, both distally and spirally corded forms known, with the latter the more common; umbilical sinus, circular, usually filled. Chambers globose, inflated, tending to less flat and somewhat convex, successive whorls strongly overlapping and narrow radially flat or only slightly convex, usually flat but sometimes low domes on last whorl increasing only slowly in size as added. Very rare specimens have last chamber smaller than penultimate chamber, formed on one side of last and always smooth. Early chambers and sutures on land area often obscured by surface ornamentation. The later sutures straight or slightly curved, depressed; sutures on umbilical side straight, depressed. Surface of test ornamented with small discontinue ribs, median arranged

Scientific Ocean Drilling

Taxa Timeline

Geological Timeline (Janet to Holocene)

Highlighted Taxon: *Margheritina undulata* (100µm Holotype)

Filter: Highlight

Scientific Ocean Drilling

Taxa Pivot

Taxon Working Group

Search...

Genus

Species

Accessory\_apertures

Sort: Accessory\_apertures

- None: 46
- Umbilical: 15
- Intalaminial: 10
- Relict: 27
- N/A: 24
- Infalaminial: 13
- Sutural: 6

Aperture

Grid of taxonomic data points (Anaticella to Biticella, Eohastigerina to Falstocrana, Globigerinelloides to Hastigerinoides, Hebergella to Laeviheterohelix, Parathalmannella to Pseudothalmannella, Rugoglobigerina to Schackonia, Thalmannella to Ticella, Whiteella)

# We can use & present 3rd party data

Unity Search Simple Box Lith Facets Visko Janus Janus Matrix

Vents ▾  
13 N Ridge Site  
ABE  
AHA Field  
**AMAR**  
Aden  
Aden New Century Mountains  
Ahyi  
Alice Springs Field  
Animal Farm  
Ashadze  
Ashadze 2

Search from vent(s) X Km  
100

This page allows searches for ocean drilling sites and related information based on ridge vents. Data from the InterRidge Global Database of Active Submarine Hydrothermal Vent Fields (the "InterRidge Vents Database") is used to allow drill site discovery. Also, links to IMLGS data is also provided.

Results

**data.oceandrilling.org SPARQL**  
**InterRidge Vents SPARQL**  
**IMLGS SPARQL**

<http://data.oceandrilling.org/codices/lsh/49>  
<http://data.oceandrilling.org/codices/lsh/49>  
<http://data.oceandrilling.org/codices/lsh/49/412> at a distance of 47.1777 Km  
<http://data.oceandrilling.org/codices/lsh/49/411> at a distance of 48.8393 Km  
<http://data.oceandrilling.org/codices/lsh/49/411/A> at a distance of 48.8393 Km  
<http://data.oceandrilling.org/codices/lsh/37/333> at a distance of 50.7856 Km  
<http://data.oceandrilling.org/codices/lsh/37/333/A> at a distance of 50.7856 Km  
<http://data.oceandrilling.org/codices/lsh/37/332/B> at a distance of 55.2069 Km  
<http://data.oceandrilling.org/codices/lsh/37/332/C> at a distance of 55.2069 Km  
<http://data.oceandrilling.org/codices/lsh/37/332> at a distance of 55.2069 Km  
<http://data.oceandrilling.org/codices/lsh/37/332/A> at a distance of 55.2069 Km  
<http://data.oceandrilling.org/codices/lsh/37/332/D> at a distance of 55.2069 Km  
<http://data.oceandrilling.org/codices/lsh/37/334> at a distance of 99.5513 Km

Card section

Map Satellite

Greenland  
Iceland  
Finland  
Sweden  
Norway  
United Kingdom  
Poland  
Germany  
France  
Spain  
Italy

North Atlantic

Google  
Map data ©2013 MapLink - Terms of Use

InterRidge URI: <http://irvents-d7.who.edu/content/amar>

General info on exp: 49  
General info on exp: 37

IMLGS: Glomar Challenger 49 Link(s)  
IMLGS: Glomar Challenger 37 Link(s)

Exploit web architecture:  
proxy, errors, CORS.. etc.

Rules used for IMLGS call

# Combining structured & unstructured data

Polyglot data structure:

Connect Solr (enhance with structured data in web pages) with Graph Data

Known connections allow us to relate URL's in Solr to URI's in Virtuoso via "relationship graph"

**STEP 1 : search on "metamorphic rock"**

**2) results in data from SOLR (which indexes sites I pick)...**

**Step 3:**  
The app uses the URLs from SOLR (like the one from NGDC) and a graph I made that associates NGDC URL's to my resource URI's. So I get pointers to things like the expedition URI for leg 72.

**step 4)**  
make the whole system SMARTER :)

**I have a couple ideas...**

**Scientific Ocean Drilling**

Search Sandbox

127.0.0.1:3030/Users/dfils/src/go/src/data.oceandrilling.org/codices/static\_labs/dart/crossref.html#metamorphic rock

metamorphic rock

Solr results:

- NOAA/NGDC/WDC for MGG, Boulder-Core Data from the Deep Sea Drilling Project (DSDP) <http://www.ngdc.noaa.gov/mgg/geology/dsdp/data/72/516F/index.htm>
- NOAA/NGDC/WDC for MGG, Boulder-Core Data from the Deep Sea Drilling Project (DSDP) <http://www.ngdc.noaa.gov/mgg/geology/dsdp/data/>
- NOAA/NGDC/WDC for MGG, Boulder-Core Data from the Deep Sea Drilling Project (DSDP) <http://www.ngdc.noaa.gov/mgg/geology/dsdp/data/59/450/1/index.htm>
- NOAA/NGDC/WDC for MGG, Boulder-Core Data from the Deep Sea Drilling Project (DSDP) <http://www.ngdc.noaa.gov/mgg/geology/dsdp/data/6/53/ndex.htm>
- NOAA/NGDC/WDC for MGG, Boulder-Core Data from the Deep Sea Drilling Project (DSDP) <http://www.ngdc.noaa.gov/mgg/geology/dsdp/data/1/1/100/1/index.htm>
- Proc. IODP, 301, Site U1301 <http://publications.iodp.org/proceedings/301/106/106.htm>
- IODP Expedition 305 Preliminary Report [http://publications.iodp.org/preliminary\\_report/305/](http://publications.iodp.org/preliminary_report/305/)
- Marine Geology and Geophysics Data at the National Geophysical Data Center | ngdc.noaa.gov <http://www.ngdc.noaa.gov/mgg/mggd.html>
- NOAA/NGDC/WDC for MGG, Boulder-Core Data from the Deep Sea Drilling Project (DSDP) <http://www.ngdc.noaa.gov/mgg/geology/dsdp/dsdpdv2.htm>
- Proc. IODP, Expedition 301, Contents <http://publications.iodp.org/proceedings/301/301toc.htm>

Graph Relations

- Resource: <http://data.oceandrilling.org/codices/ish/72>
- Resource: <http://data.oceandrilling.org/codices/ish/72/516>
- Resource: <http://data.oceandrilling.org/codices/seas/1/1914>
- Resource: <http://data.oceandrilling.org/codices/ish/6>
- Resource: <http://data.oceandrilling.org/codices/ish/6/53/A>
- Resource: <http://data.oceandrilling.org/codices/ish/6/53/B>
- Resource: <http://data.oceandrilling.org/codices/seas/1/4300>
- Resource: <http://data.oceandrilling.org/codices/ish/6/44>
- Resource: <http://data.oceandrilling.org/codices/ish/6/45>
- Resource: <http://data.oceandrilling.org/codices/ish/6/45/A>
- Resource: <http://data.oceandrilling.org/codices/ish/6/46>
- Resource: <http://data.oceandrilling.org/codices/ish/6/47>
- Resource: <http://data.oceandrilling.org/codices/ish/6/47/A>
- Resource: <http://data.oceandrilling.org/codices/ish/6/47/B>
- Resource: <http://data.oceandrilling.org/codices/ish/6/48>
- Resource: <http://data.oceandrilling.org/codices/ish/6/48/A>
- Resource: <http://data.oceandrilling.org/codices/ish/6/48/B>
- Resource: <http://data.oceandrilling.org/codices/ish/6/49>
- Resource: <http://data.oceandrilling.org/codices/ish/6/49/A>
- Resource: <http://data.oceandrilling.org/codices/ish/6/50>
- Resource: <http://data.oceandrilling.org/codices/ish/6/50/A>
- Resource: <http://data.oceandrilling.org/codices/ish/6/51>
- Resource: <http://data.oceandrilling.org/codices/ish/6/51/A>

NOAA/NGDC/WDC for MGG, Boulder-Core Data from the Deep Sea Drilling Project (DSDP) <http://www.ngdc.noaa.gov/mgg/geology/dsdp/data/72/516F/index.htm>

NOAA NATIONAL GEOPHYSICAL DATA CENTER  
MARINE GEOLOGY AND GEOPHYSICS DIVISION  
World Data Center for Marine Geology & Geophysics, Boulder  
Seafloor Series Volume 1

Core Data from the Deep Sea Drilling Project

Lea 72, hole 516F

Physiographic feature: rise  
Total penetration (m): 1271  
# sediment cores: 28  
Oldest sediment core: 124  
Oldest sediment age: Lower Senonian  
Oldest sediment description: calcarenite and breccia  
Type of crust: oceanic  
Depth to basement (m): 1253  
# Rock cores: 3  
Rock description: basalt  
Other holes at this site: [516](#) [516A](#) [516B](#) [516C](#) [516D](#) [516E](#)  
Adjacent sites : [22](#) [357](#)

Data types available:

age profile	<a href="#">original</a>	<a href="#">delimited</a>	
carbon/carbonate	<a href="#">original</a>	<a href="#">delimited</a>	
core depth recovery	<a href="#">original</a>	<a href="#">delimited</a>	
density-porosity	<a href="#">original</a>	<a href="#">delimited</a>	
grain size	<a href="#">original</a>	<a href="#">delimited</a>	
gamma ray attenuation porosity evaluator	<a href="#">original</a>	<a href="#">delimited</a>	
igneous/metamorphic rock descriptions	<a href="#">original</a>	<a href="#">delimited</a>	<a href="#">browse</a>
igneous/metamorphic rock paleomagnetism	<a href="#">original</a>	<a href="#">delimited</a>	<a href="#">browse</a>
igneous/metamorphic rock major-element chemistry	<a href="#">original</a>	<a href="#">delimited</a>	<a href="#">browse</a>
igneous/metamorphic rock minor-element chemistry	<a href="#">original</a>	<a href="#">delimited</a>	<a href="#">browse</a>
planktonic foraminifera	<a href="#">original</a>	<a href="#">delimited</a>	<a href="#">browse</a>
SCREEN descriptions	<a href="#">original</a>	<a href="#">delimited</a>	<a href="#">browse</a>
site summary information	<a href="#">original</a>	<a href="#">delimited</a>	
smearslide descriptions	<a href="#">original</a>	<a href="#">delimited</a>	<a href="#">browse</a>
sonic velocity	<a href="#">original</a>	<a href="#">delimited</a>	
visual descriptions	<a href="#">original</a>	<a href="#">delimited</a>	<a href="#">browse</a>
water chemistry (interstitial)	<a href="#">original</a>	<a href="#">delimited</a>	



# How we approached this

LOTS of entry points into the data!

Bulk data access (Tools like R, iPython, Corewall, Google Earth, GeoMapApp (someday?), etc as well as Web)

Triple Server is THE point of departure, but polyglot persistence behind the scene

Decouple the stack! As much as you link data.. don't link the code base and web architecture!

Go and Dart(JS) as the code base.

# Practices in different spaces

RDF space: (303's, negotiation, SKOS, SPARQL, as many community vocabularies as we possibly can.)

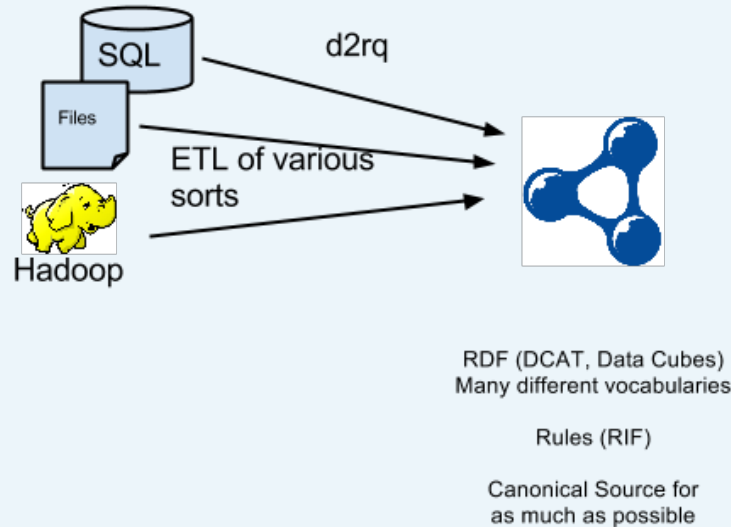
HTML space: (RDFa, schema.org with things like DCAT and other vocabularies)

Federation space: (SPARQL 1.1, **resolve on subject** and less so predicate) VOID perhaps?

Other: OpenRefine reconcile API, RIF (and Rules in general)

Getting along with REST: JSON-STAT, JSON-LD and Linked Data API and DCAT and RDF Data-Cubes

# Data Flow Landscape



Query

SPARQL endpoint

API's

OpenRefine Reconciliation API

REST: OpenSearch API

Linked Data API

Interop

REST: Interop via JSON-  
STAT / JSON-LD

Common URI's

Vocabulary space

Web

WWW: Linked Open Data

WWW: RDFa and schema.org

LOD linking: Freebase and more

# Goals and Future

Glue the data together even more (and more use of vocabulary aspects)

Flesh out Discovery (and use)

Reconcile and Disambiguate "things"