CONSORTIUM FOR
Ocean Leadership

Data Discovery and Integration
An exploration of concepts for IODP

Douglas Fils
dfils@oceanleadership.org

Ocean Leadership

On efforts at 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

Ocean Leadership

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...

Ocean Leadership

How we approached this / What we did

On Federation:

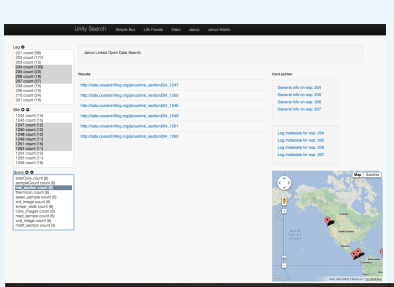
- Linked Open Data (LOD) & vocabularies as a basis to explore approaches
- RDF + SPARQL
- 5 star data
- Query vs API

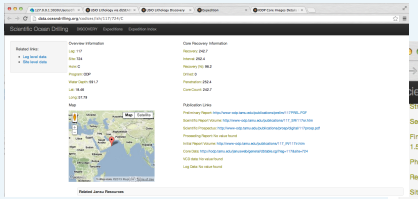
On Discovery:

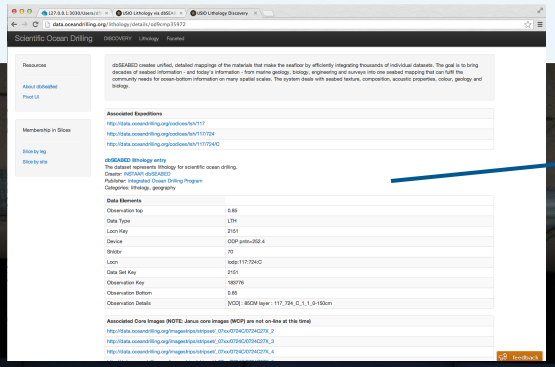
- Differentiate "discovery" from "search"
- Encourage 3rd parties to do the discovery and inspection of data
- Enable better discovery via Google et al
- Work with other related providers to link in URI and vocabulary space

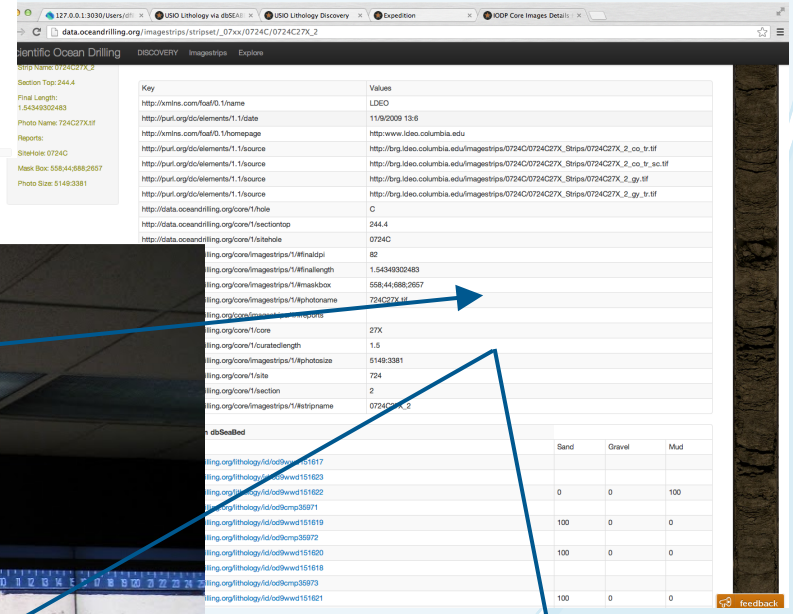
Ocean Leadership

Federation (provides discovery and exploration)

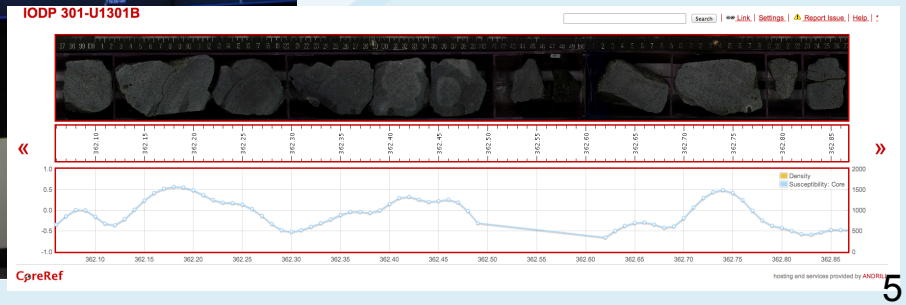




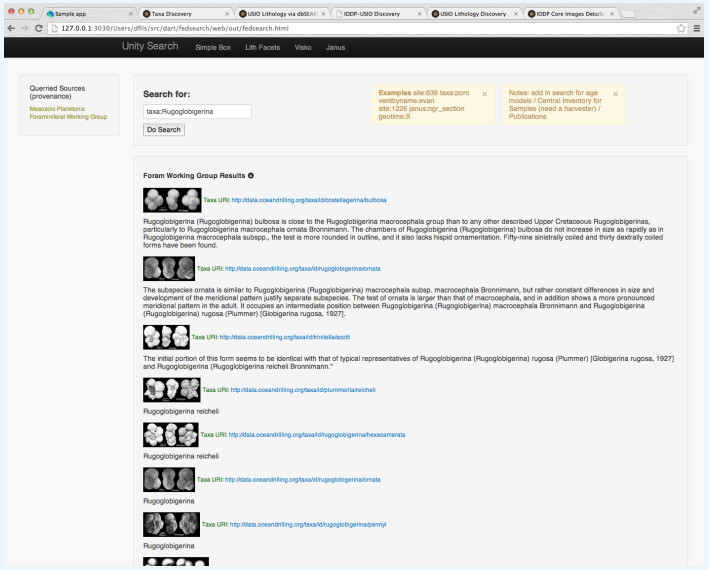


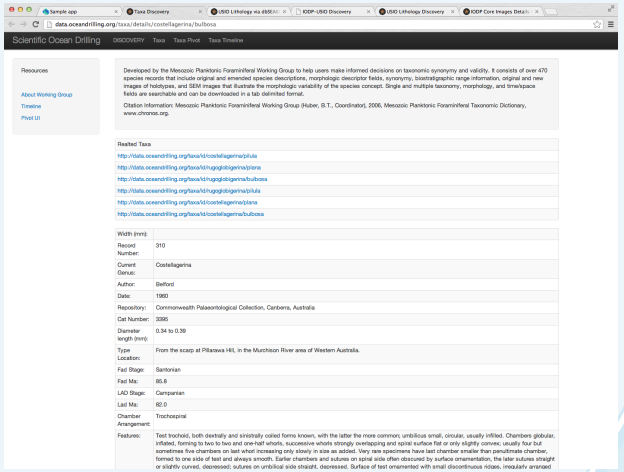


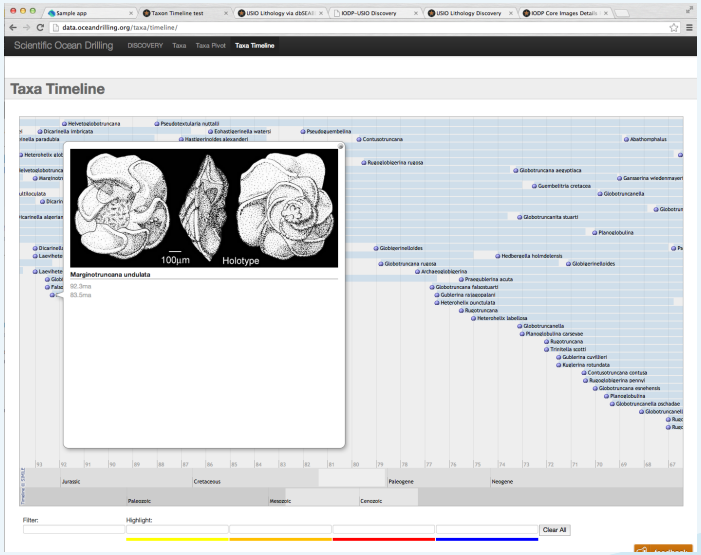
3rd party tools

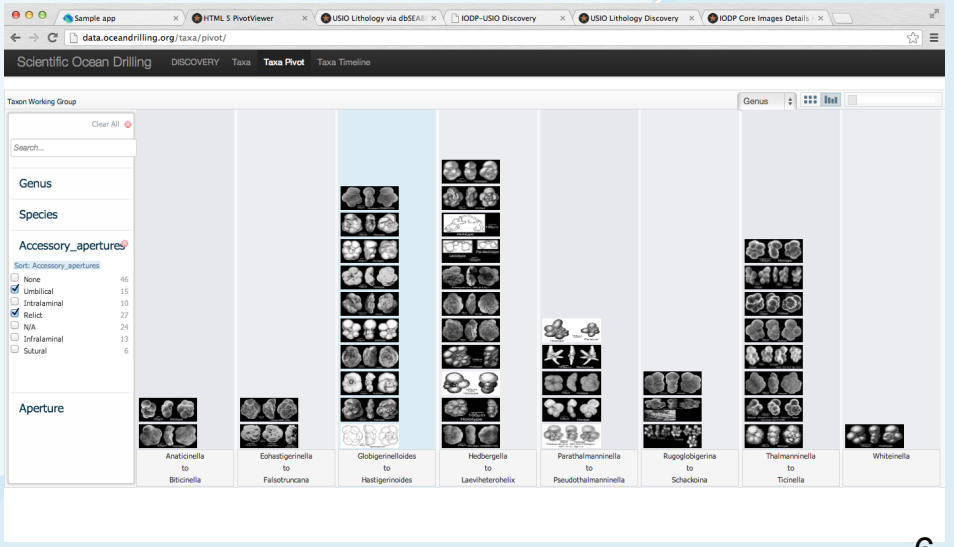


Multiple viewing options

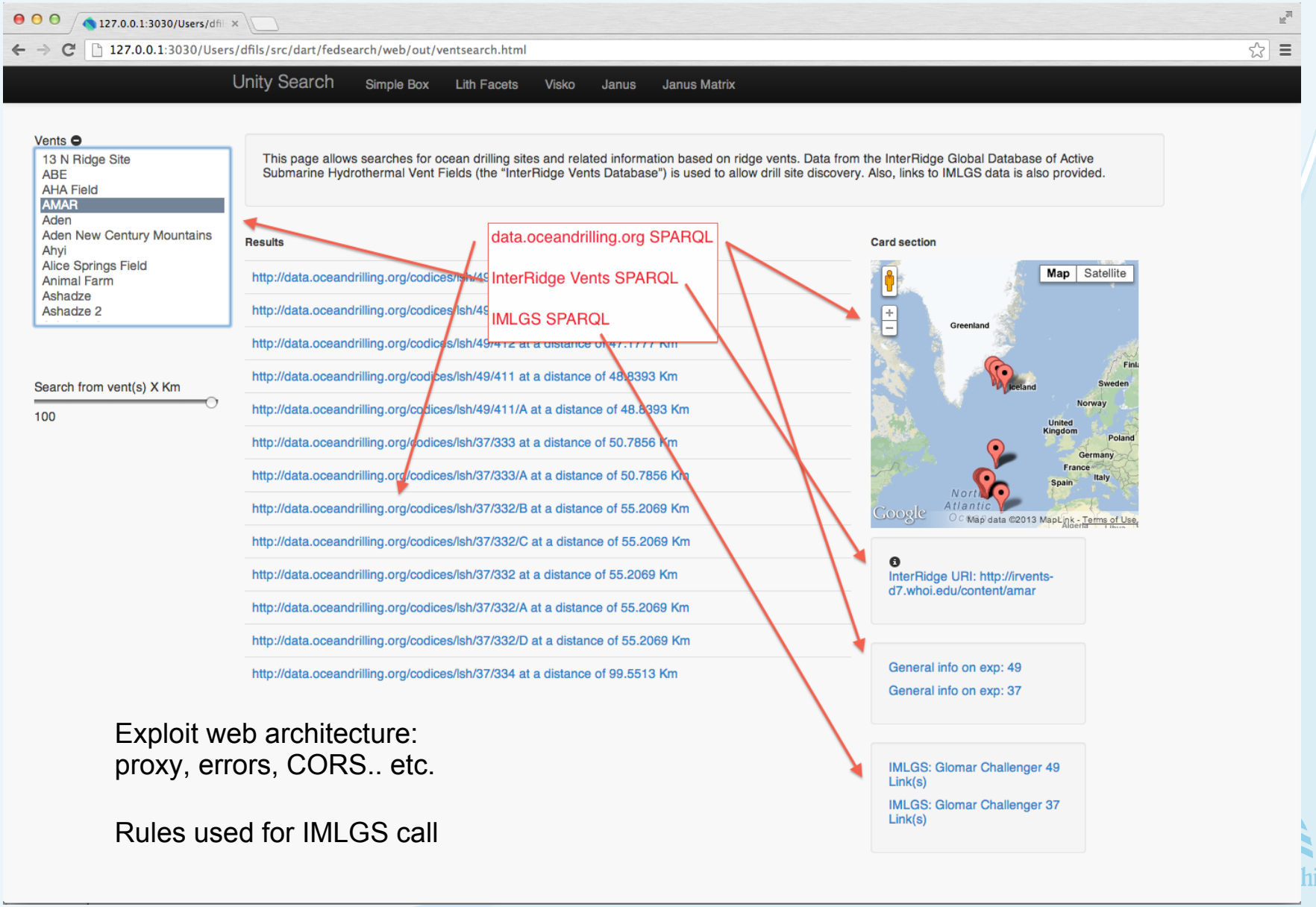








We can use & present 3rd party data



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

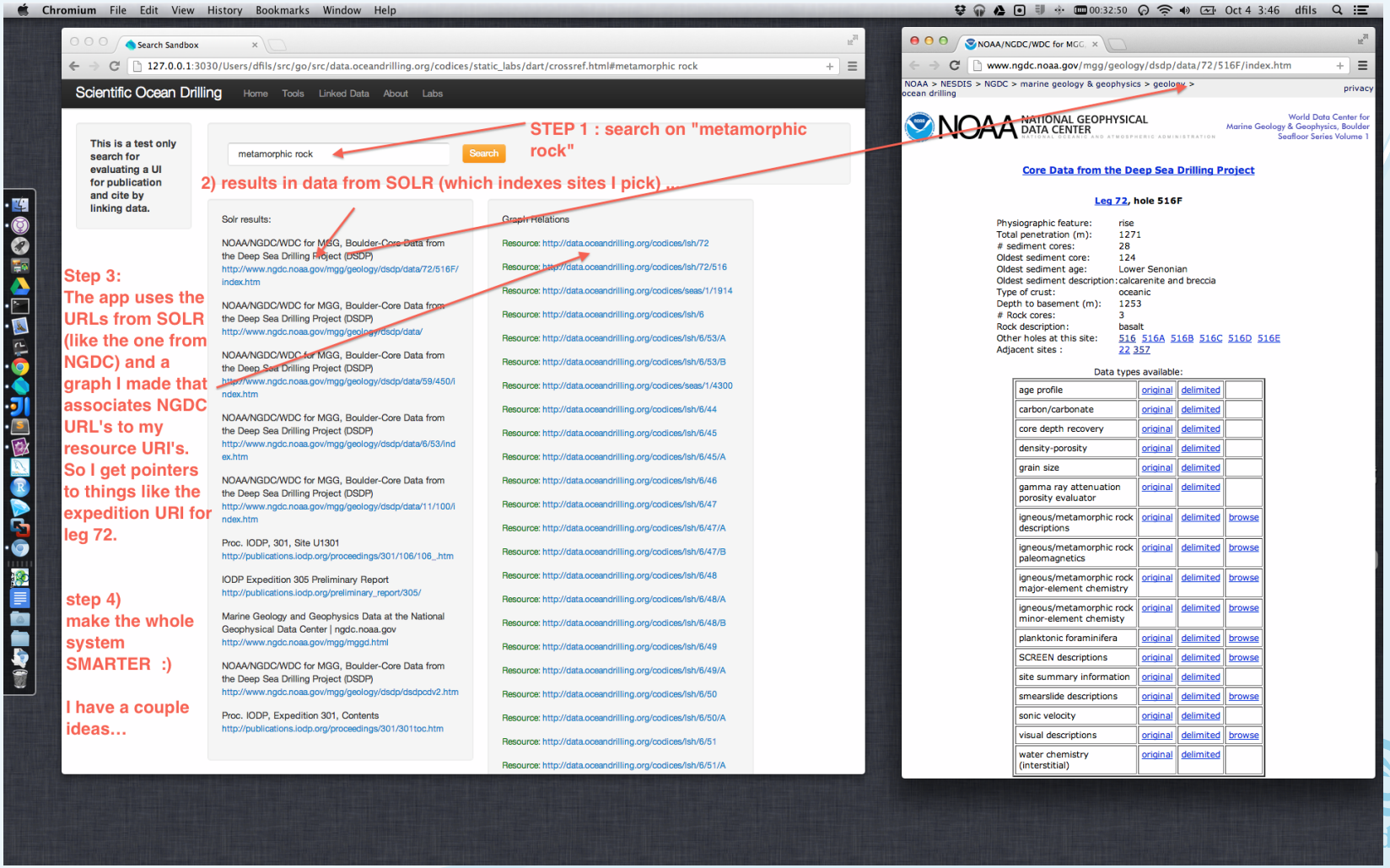
Rules used for IMLGS call

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"



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.

Ocean Leadership

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)

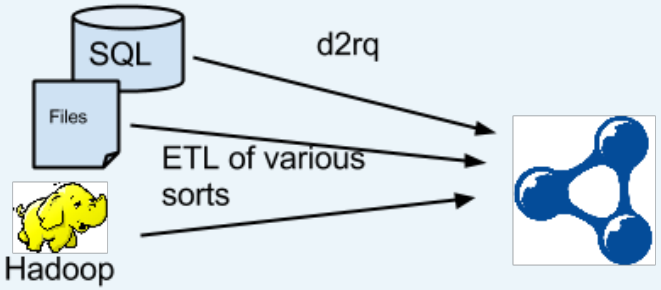
Federation space: (SPARQL 1.1, **resolve on subject** and less so predicate) VOID perhaps?

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

Ocean Leadership

Data Flow Landscape



RDF (DCAT, Data Cubes)
Many different vocabularies

Rules (RIF)

Canonical Source for as much as possible

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

Ocean Leadership

Goals and Future

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

Flesh out Discovery (and use)

Reconcile and Disambiguate "things"

Ocean Leadership