Interoperability in Science Data: Stories from the Trenches

Karen Stocks
University of California San Diego

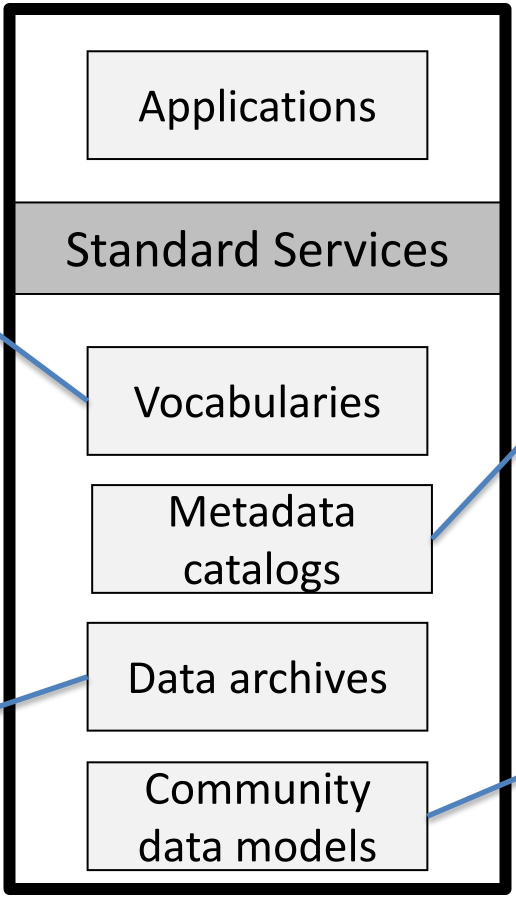
Open Data for Open Science – Data Interoperability
Microsoft eScience Workshop 2012

Louisville
NSF
M
NSF
L

Interoperability Case Studies

- Ocean Observatories Initiative
- Rolling Deck to Repository
  - internal operations
  - external interactions

Earth Cube Cross-Domain Interoperability Framework



Data available in compatible semantics: ontologies, controlled vocabularies

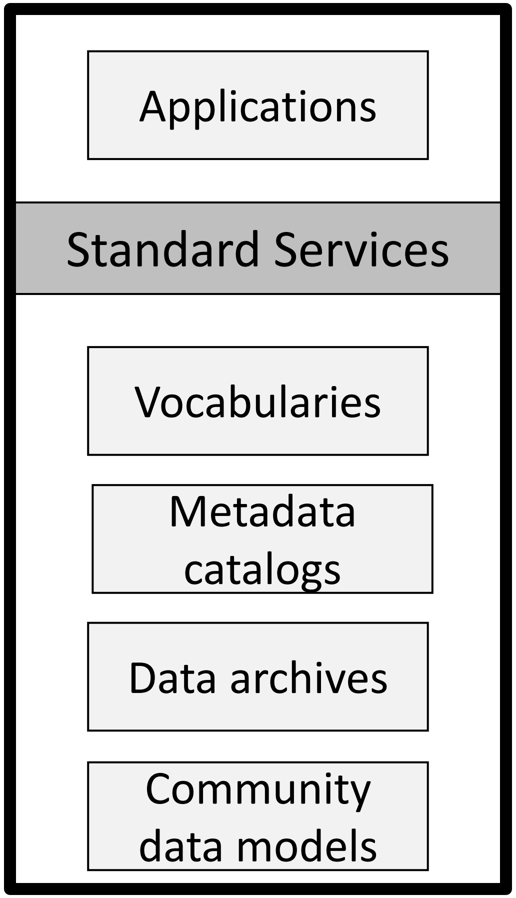
Find and retrieve domain resources: files and file collections, services, documents - by thematic category, type, location

Data available via standard service interfaces (e.g. OGC WFS, SOS) but different information models

Compatibility at the level of domain information models and databases

Interop Framework - shorthand

- – Vocabularies
- – Metadata
- – Data Access
- – Data Models



Applications

Standard Services

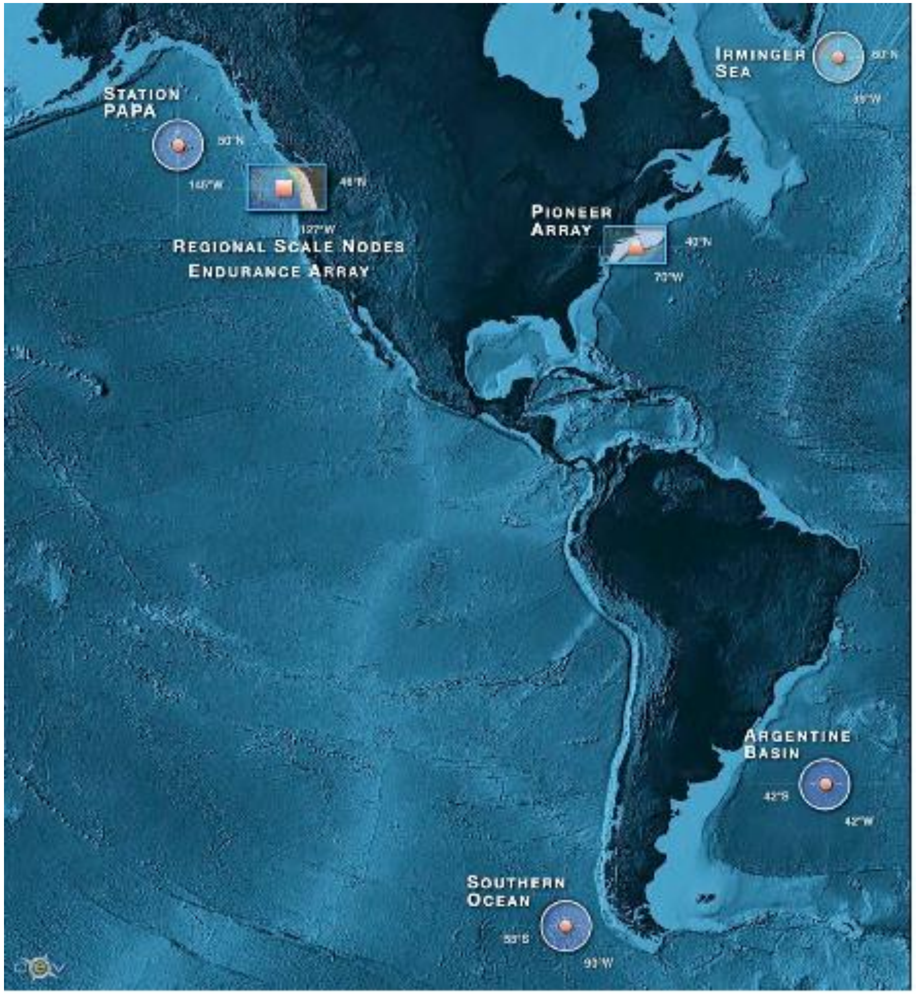
Vocabularies

Metadata catalogs

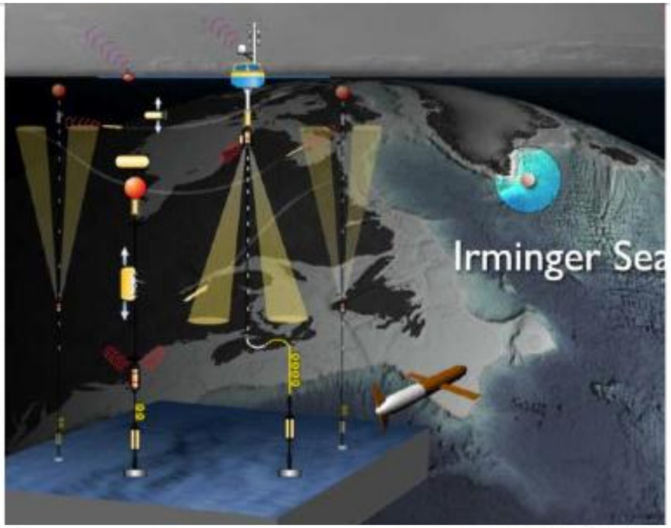
Data archives

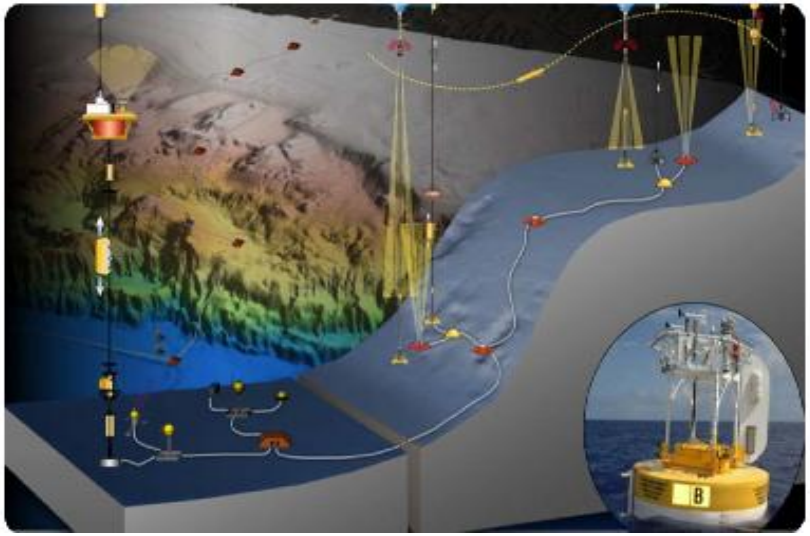
Community data models

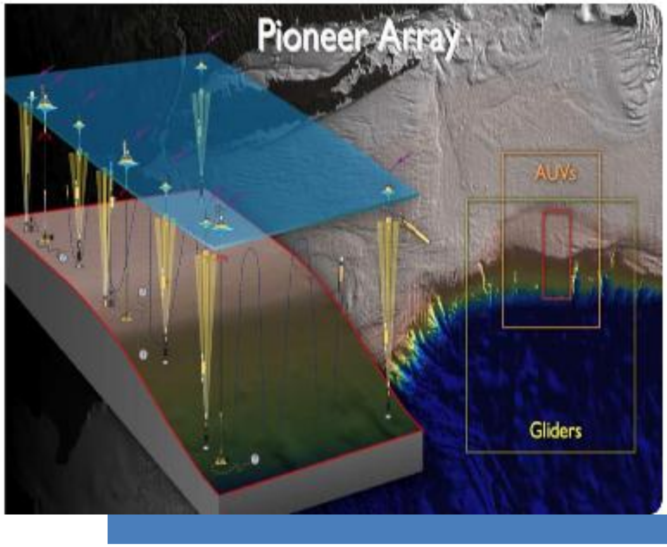
The Ocean Observatories Initiative (OOI): Instrumenting the Oceans

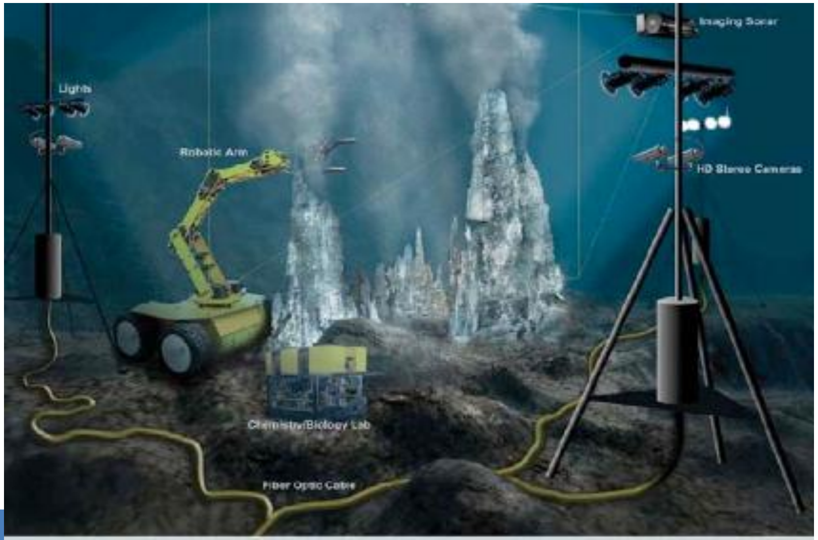


Long-term, in-situ instrumentation

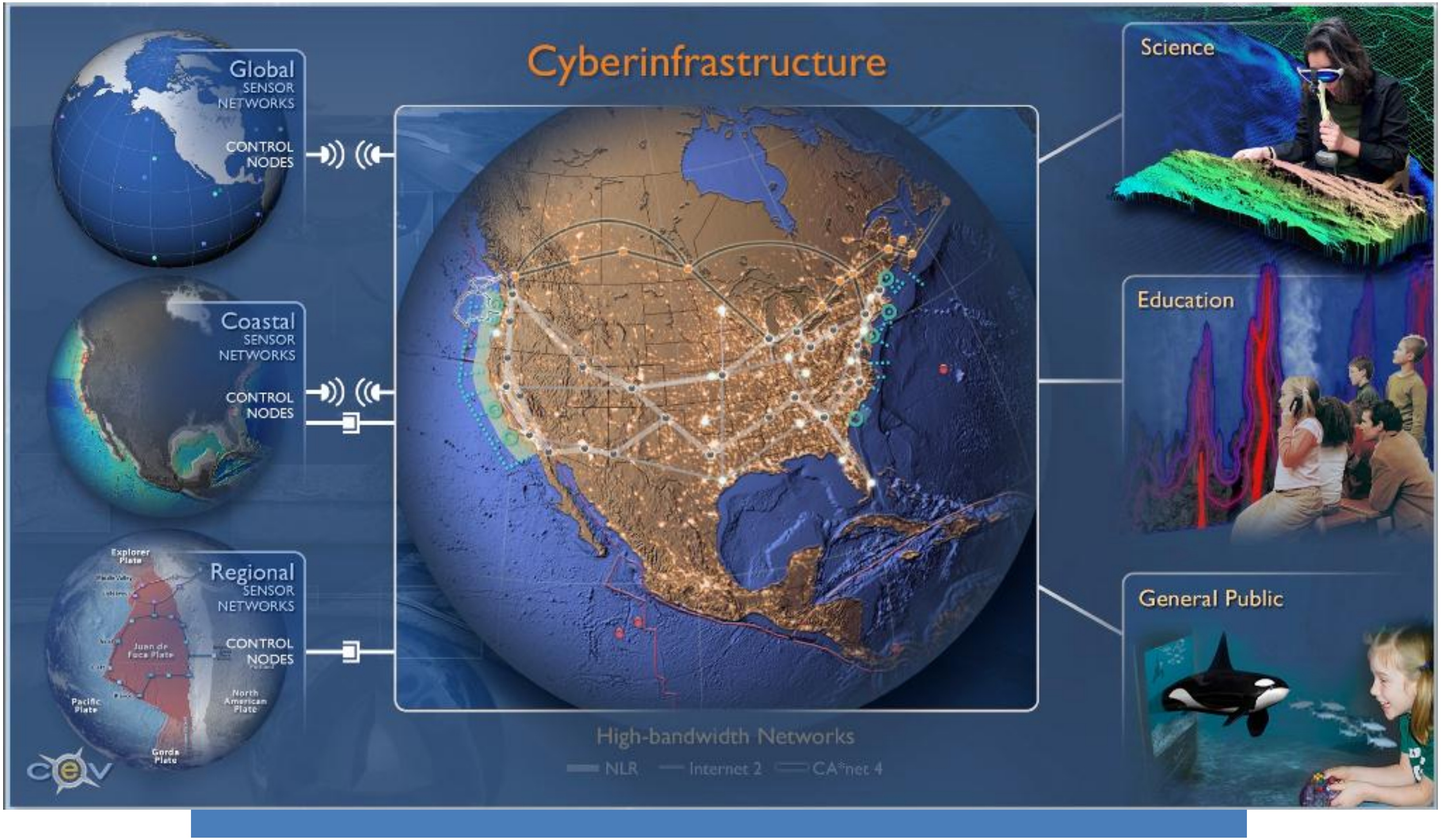








Cyberinfrastructure: linking the marine infrastructure to science and user



OOI Challenges

Vocabulary and Data Format

OOI will have~50 instrument types

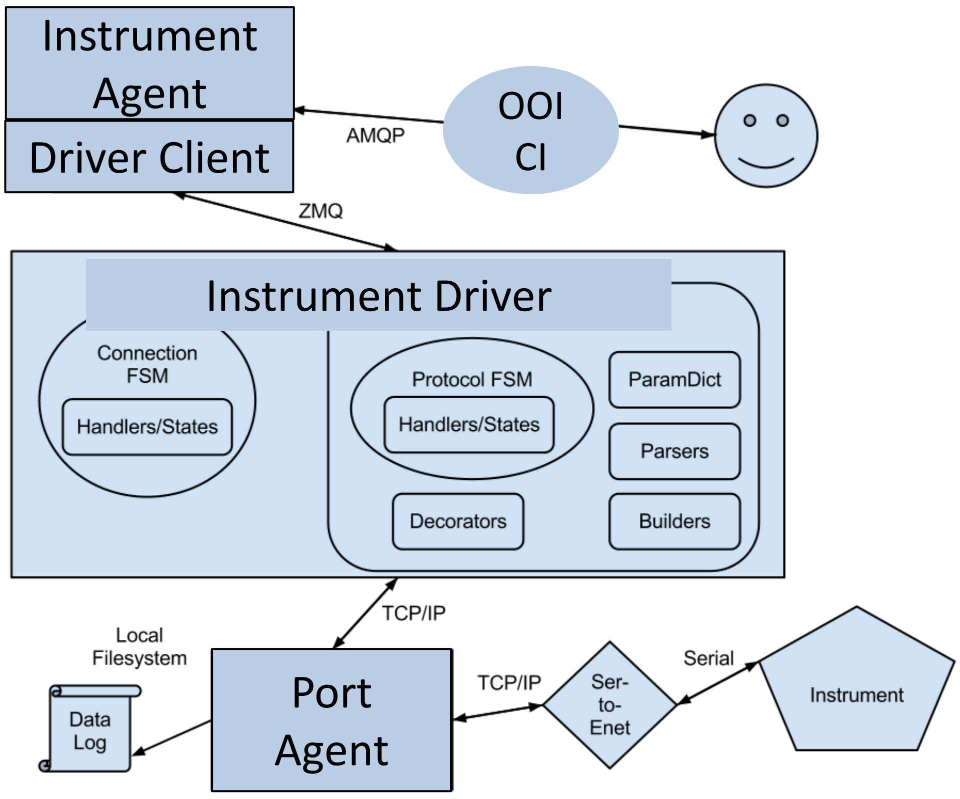
The Instrument Operator on shore should not have to know how to task 50 different instruments; there needs to be a single set of basic commands (power on, take a reading, start autosampling, etc.)

- The heterogeneous data must flow into a common data model

OOI Challenges

Vocabulary and Data Format

Solution: write drivers and agents for each instrument class/model, and Data Processing algorithms for data ingestion



K. Stocks, Microsoft eScience Workshop 2012

OOI Challenges
Vocabulary and Data Format

Assessment

✔ Effective: instrument can be tasked and data ingested

✗ Scalable: new (different) instruments require new effort

✗ Interoperability: no gains outside of OOI

OOI Challenges

Vocabulary and Data Format

A better (partial) solution: Partner with instrument manufacturers to further develop and adopt a framework for describing sensor data provenance, building off of Open Geospatial Consortium SensorML, to allow standards-based, machine-harvestable encodings (Janet Fredericks, WHOI) rvdata.us

Rolling Deck to Repository (R2R)

Managing underway data from research vessels

NSF

Lamont-Doherty Earth Observatory
COLUMBIA UNIVERSITY | EARTH INSTITUTE

SCRIPPS INSTITUTION OF OCEANOGRAPHY
UC SAN DIEGO

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

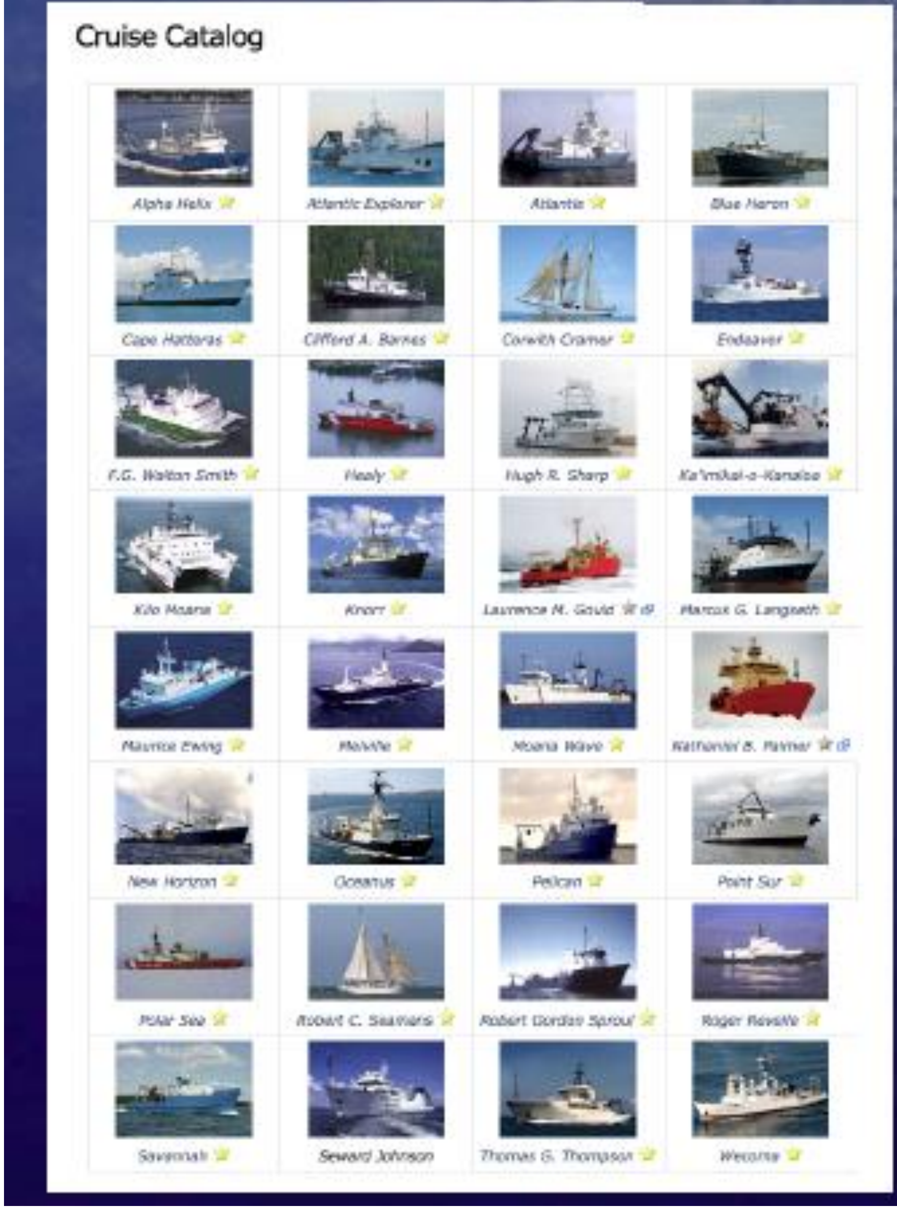
WOODS HOLE OCEANOGRAPHIC INSTITUTION
1930

SAMOS

FLORIDA STATE UNIVERSITY
1851

K. Stocks, Microsoft eScience Workshop 2012

Academic Fleet



Cruise Catalog

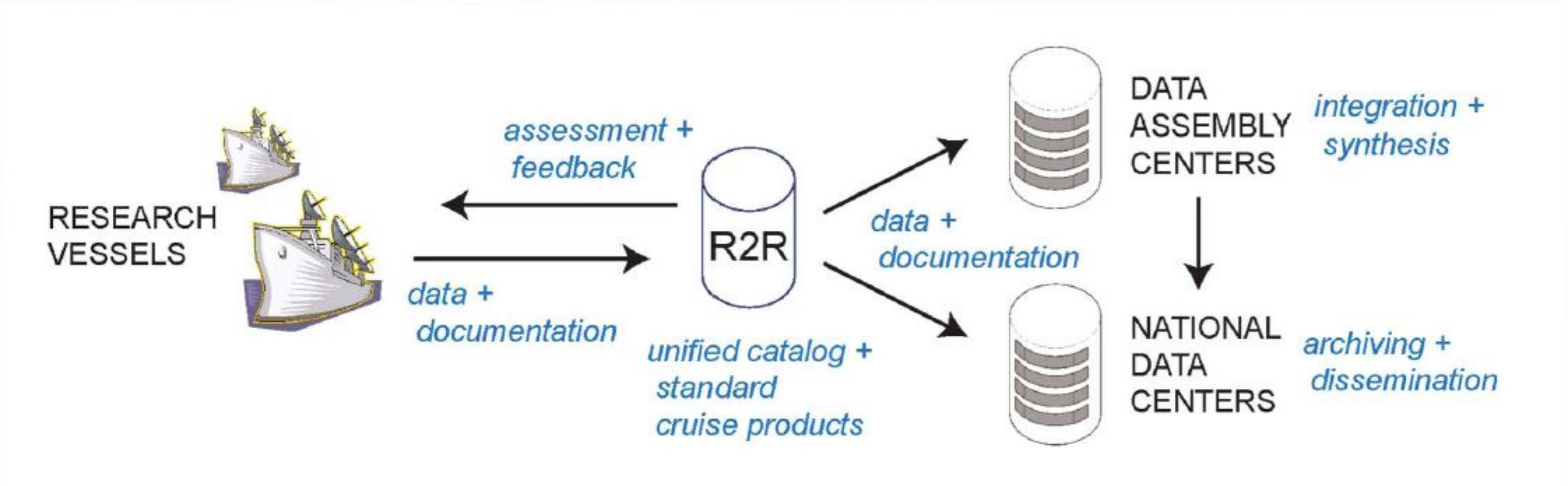
Alpha Helix	Atlantic Explorer	Atlantis	Blue Heron
Cape Hatteras	Clifford A. Barnes	Corwith Cramer	Endeavor
F.G. Walton Smith	Healy	Hugh R. Sharp	Ka'imikai-o-Kanaloa
Kilo Moana	Knorr	Laurence M. Gould	Marcus G. Langseth
Maurice Ewing	Melville	Moana Wave	Nathaniel B. Palmer
New Horizon	Oceanus	Pelican	Point Sur
Polar Sea	Robert C. Seamans	Robert Gordon Sproul	Roger Revelle
Savannah	Seward Johnson	Thomas G. Thompson	Wecoma

- 30+ vessels in active or recent service
- 100's of cruises/year
- 1000's of datasets/year

Joined R2R

Oct. 2008	*Langseth*
	(Healy)
	Kilo Moana
	Melville
	Revelle
2009	*Thompson*
	Sharp
	Atlantis
	Knorr
	Oceanus
	(Ka'imikai)
	Barnes
	Walton Smith
2010	*Point Sur*
	Wecoma
	(Cramer)
	(Seamans)
	Endeavor
	New Horizon
	Sproul
	(Polar Sea)
	Pelican
	Savannah
	Explorer
	Hatteras
★Oct. 2010	*Blue Heron*

R2R Goals

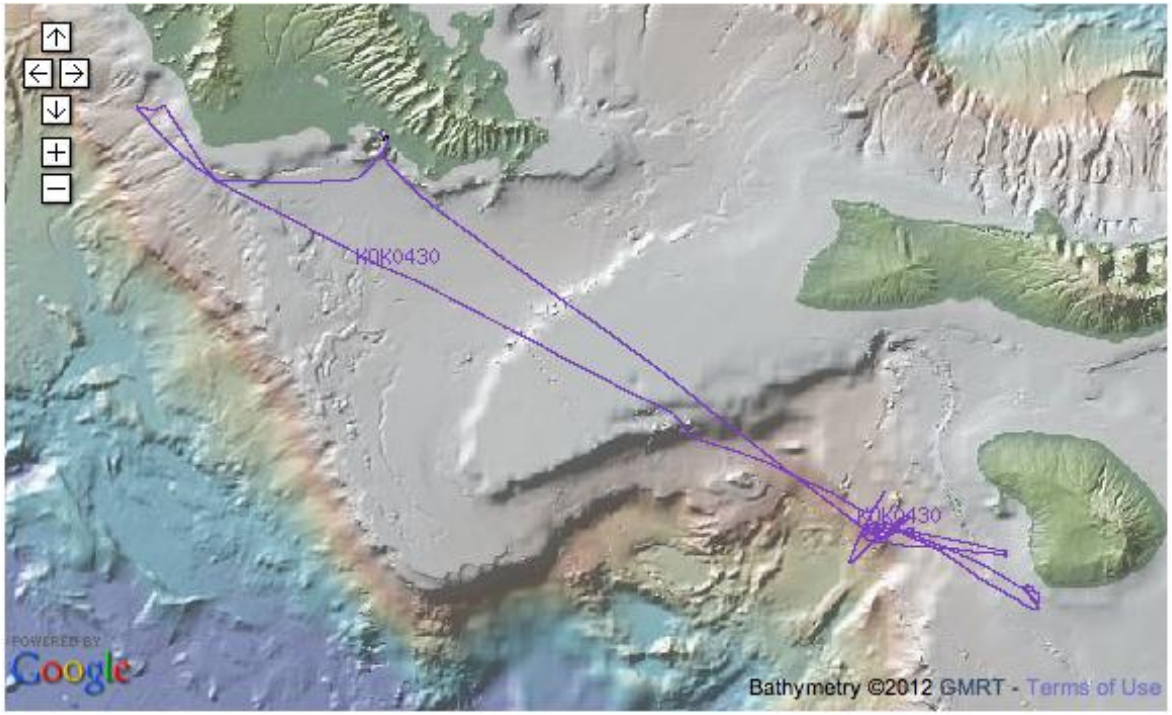


For the U.S. Academic Oceanographic Research Fleet:

- Migrate all routine “underway” data to long-term repositories
- Create catalog of cruises and standard products
- Assess data quality and provide timely feedback to vessels

R2R Challenges
1. Metadata

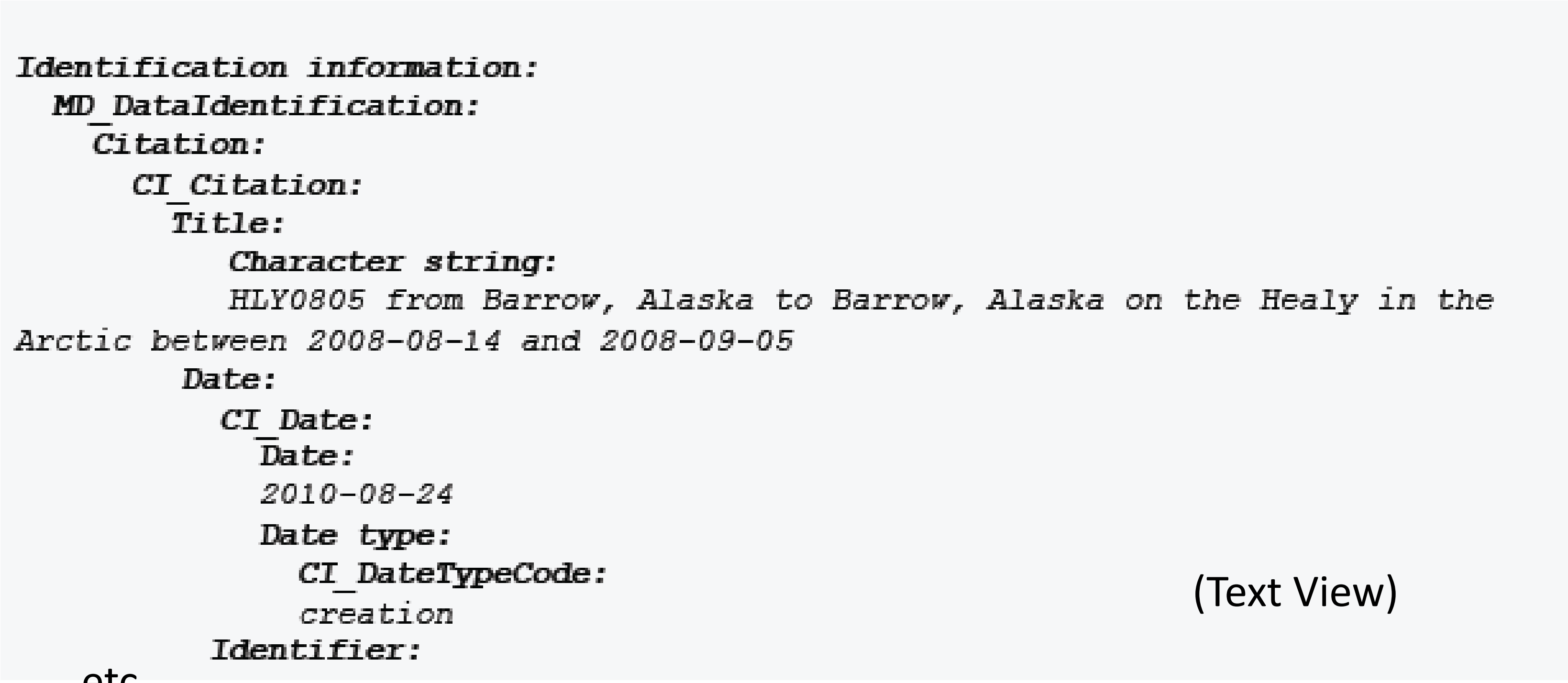
Problem: R2R collects critical discovery and access metadata about a "new" level of granularity: a cruise



R2R Challenges

1. Metadata

Solution: Work with the National Geophysical Data Center to create a new ISO-19115 compliant cruise-level metadata standard.



```
Identification information:
  MD_DataIdentification:
    Citation:
      CI_Citation:
        Title:
           Character string:
           HLY0805 from Barrow, Alaska to Barrow, Alaska on the Healy in the
Arctic between 2008-08-14 and 2008-09-05
         Date:
           CI_Date:
             Date:
             2010-08-24
             Date type:
               CI_DateTypeCode:
               creation
          Identifier:
```

(Text View)

…etc.

R2R Challenges

1. Metadata

Assessment:

✔ Effective: substantial increase in discoverability and usability of data

✔ Scalable: after initial time investment, now auto-generated for new cruises

✔ Interoperability: builds off of existing generic ISO standard, creates new framework for others

R2R Challenges

2. Data Format

Problem: large heterogeneity in formats for data coming from independent ship operators…even when the same instrument is being used.

Solution: parser for each data format variant; transformation to standard format for certain data types

R2R Challenges

2. Data Format

Assessment

✔ Effective: allows data to be accessed

✗ Scalable: every new format requires same level of new effort.

~ Interoperability: approach creates no interoperability gains beyond reformatting done by R2R

R2R Challenges

2. Data Format

A better solution: community adoption of format standards

BUT…for some data types, these already exist, and are simply not being used, mainly because of cost of initial change

➔This needs a human/resource solution, not a technology solution

R2R Challenges, 3

Thule
Barrow
Nuuk (Godthab)
Groennedal
Prince Rupert
ZHNG09RR
ZHNG08RR
Duluth
KNOX18RR
Southampton
Coos Bay
Redwood City
Baltimore
Gibraltar

Oceanographic Vessel Data are global…and should be globally accessible. Emerging Ocean Data Interoperability Platform is addressing

Rio De Janeiro
Brisbane
Fremantle
RR0903
Isla De Pascua
AMAT02RR
Montevideo
Puerto Montt
ZHNG02RR
Dunedin
RR0902
KNOX14RR
RR0901
Punta Arenas
KNOX03RR
POWERED BY
Google
Bathymetry ©2012 GMRT - Terms of Use

R2R Challenges

3. Vocabulary

Problem: similar concepts, such as countries, vessels, instruments, and datasets, are used by many oceanographic information systems.

R2R Challenges
3. Vocabulary

Solution 1: Use existing controlled vocabularies where available

- Country (ISO)
- Cruise Type (UNOLS)
- Gazetteer - Exclusive Economic Zone (VLIZ)
- Gazetteer - Sea Area (IHO)
- Gazetteer - Undersea Feature Name (IHO)
- Language (ISO)
- Organization (IANA)
- Port (UNOLS)
- Processing Level (CODMAC)
- Sample Type (USGS)
- State (FIPS)
- Vessel (ICES)

R2R Challenges
3. Vocabulary

Solution 2: R2R and partner organizations are adopting a linked data approach to make catalog content broadly and easily accessible.

- RDF & URIs
- SPARQL endpoints
- D2RQ
- DOIs

Semantic Web

UC3EZID
long-term identifiers made easy

DataCite

Outside R2R Challenges

Vocabulary

Assessment:

? Effective: Too soon to assess use

✔ Scalable: Efficient re-use of vocabularies

✔ Interoperability: at vocabulary and Link level

A Question for Future Cross-Domain Interoperability

To what degree should we mandate global standards or allow local domain-specific protocols?

Some success in these projects with a global framework, and a local extension

- e.g. R2R Cruise metadata standard, from ISO 19115
- Instrument self-reporting provenance, built on OGC SensorML

Is this extensible? Are there better approaches?