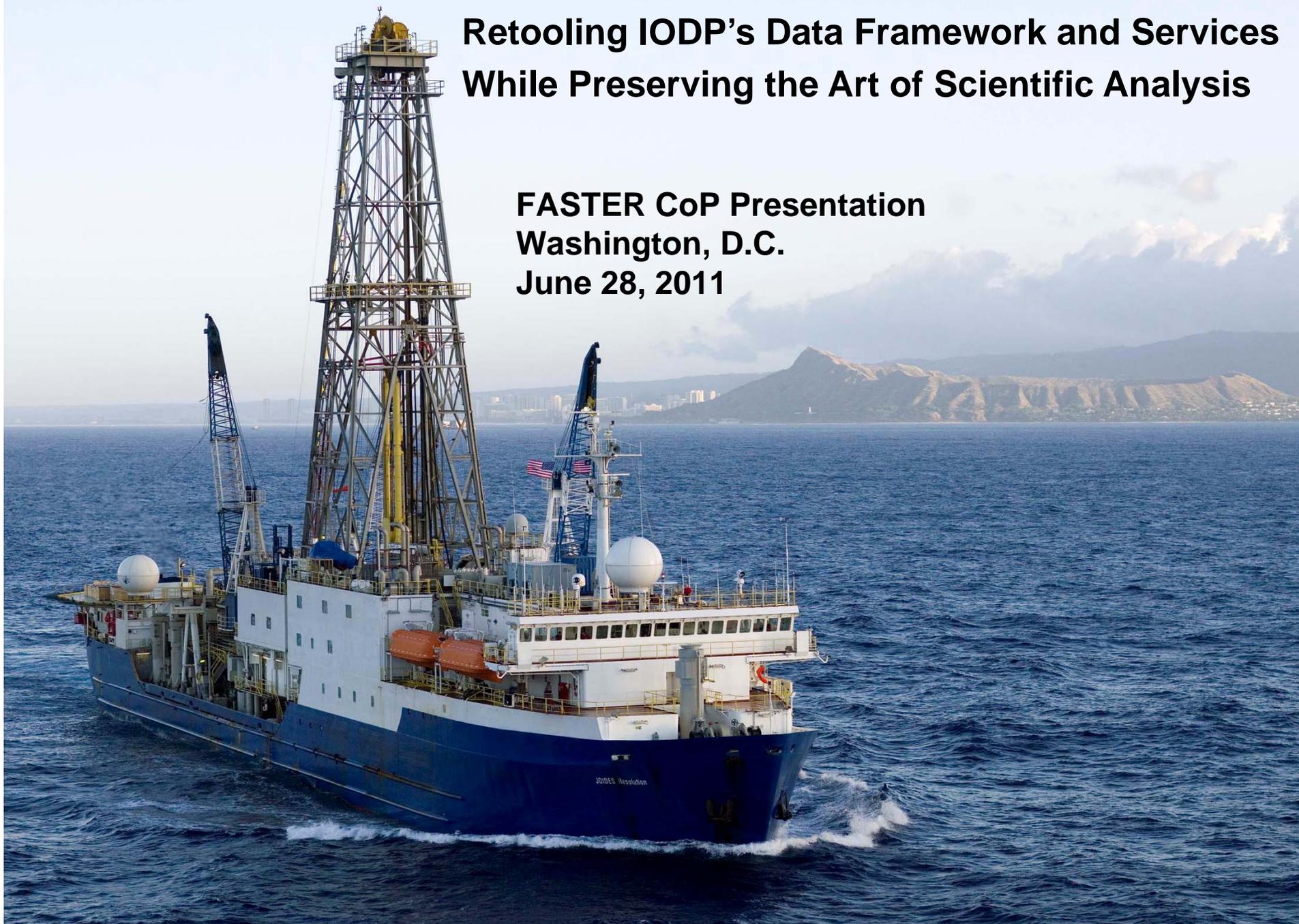


The Road to Better Data Collaboration in Geosciences: Retooling IODP's Data Framework and Services While Preserving the Art of Scientific Analysis

**FASTER CoP Presentation
Washington, D.C.
June 28, 2011**



Outline



- **What is IODP?**
- **What was our data dilemma?**
- **What was our solution?**
- **What did we learn?**
- **What is our vision?**

What is IODP?

Integrated Ocean Drilling Program

**Internationally funded (6 international entities,
24 contributing member countries)**

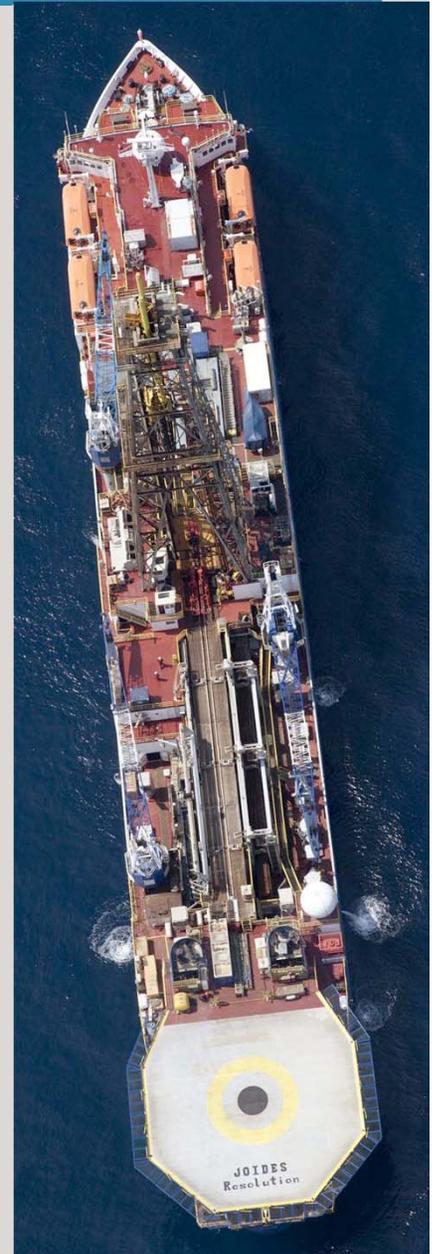
**Consortium for Ocean Leadership, Texas A&M
University, Lamont-Doherty Earth Observatory
are the US Implementing Organization**

Mission: Advancing understanding of Planet Earth
– **Sampling, analyzing, and monitoring
subseafloor environments**

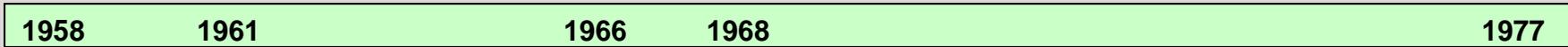
**Fully collaborative research (multinational,
multilingual, broad spectrum of computing
platforms - all customer controlled)**



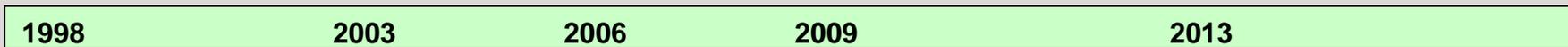
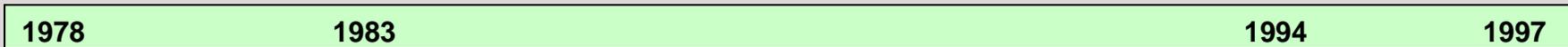
Integrated Ocean Drilling Program
United States Implementing Organization



Evolution of scientific ocean drilling



Project Mohole 



Integrated Ocean Drilling Program
United States Implementing Organization

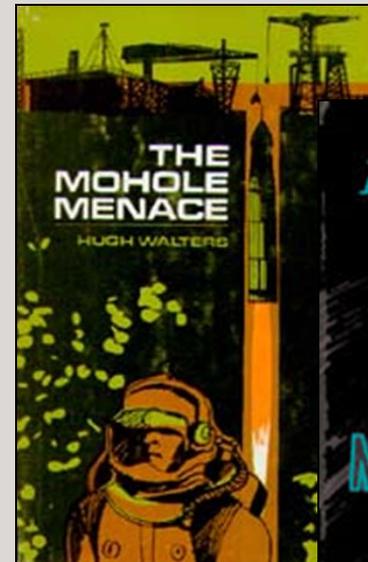
Project Mohole



CUSS 1

Phase 1 - Offshore Guadalupe Island
(Baja California, Mexico) March - April 1961

3500m water depth, five holes
183m sediments to 15 Mya
13.5m basalt



Integrated Ocean Drilling Program
United States Implementing Organization

Project Mohole



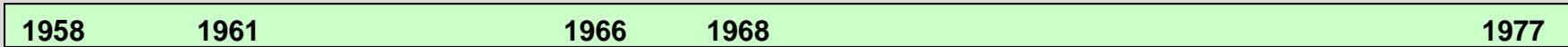
CUSS 1 is the first floating drilling rig

Life magazine correspondent
John Steinbeck



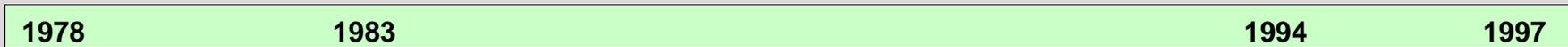
Integrated Ocean Drilling Program
United States Implementing Organization

Evolution of scientific ocean drilling



Project Mohole 

Deep Sea Drilling Project 



DSDP 



Integrated Ocean Drilling Program
United States Implementing Organization

Deep Sea Drilling Project

Glomar Challenger

NSF prime contract

International participation

By 1976, 5 international
funding partners

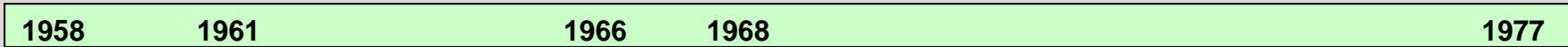
Data files compiled from paper
forms

Data files available from National Geophysical Data Center or
in tabular form in printed volumes



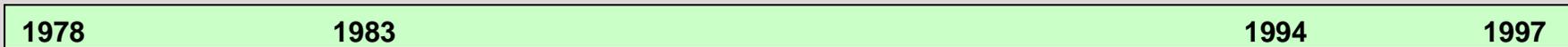
Integrated Ocean Drilling Program
United States Implementing Organization

Evolution of scientific ocean drilling

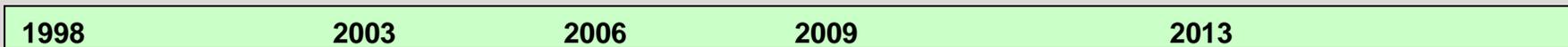


Project Mohole 

Deep Sea Drilling Project 



DSDP  **Ocean Drilling Program** 



ODP 



Integrated Ocean Drilling Program
United States Implementing Organization

Ocean Drilling Program



JOIDES Resolution

Internationally funded and staffed

1994-Began development of JANUS database

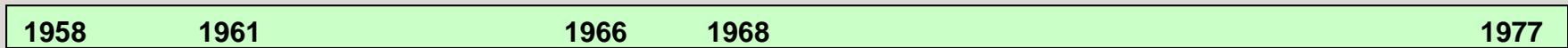
1998-Implementation of JANUS and evolution to electronic publications

Estimated <50% of data collected in database



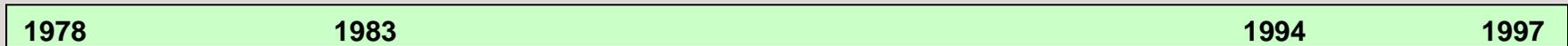
Integrated Ocean Drilling Program
United States Implementing Organization

Evolution of scientific ocean drilling

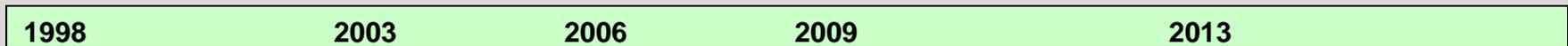


Project Mohole 

Deep Sea Drilling Project 



DSDP  **Ocean Drilling Program** 



ODP  **Integrated Ocean Drilling Program** 



Integrated Ocean Drilling Program
United States Implementing Organization

Integrated Ocean Drilling Program



Refit JOIDES Resolution -
USIO



CHIKYU –
CDEX Japan



Mission-Specific Platforms – ECORD



Integrated Ocean Drilling Program
United States Implementing Organization

Integrated Ocean Drilling Program



Phase 1 2003-2006

JANUS database

ODP data acquisition systems

SODV 2006-2009

\$115 million refit project

Major ship infrastructure changes

New instruments and data systems

Phase 2 2009-2013

Laboratory Information

Management System

New or revamped data

acquisition systems



Integrated Ocean Drilling Program
United States Implementing Organization

Evolution of scientific ocean drilling

1958 1961 1966 1968 1977

Project Mohole 

Deep Sea Drilling Project 

1978 1983 1994 1997

DSDP  **Ocean Drilling Program** 

1998 2003 2006 2009 2013

ODP  **Integrated Ocean Drilling Program**  **International Ocean Discovery Program**



Integrated Ocean Drilling Program
United States Implementing Organization

Our customers' expectations



Data gathering

**Provide tools and technology for earth materials characterization
Geography, Geology, Geochemistry, Geophysics, Geomicrobiology
Stakeholder-defined data acquisition systems
Customer-supplied data acquisition systems**

Data analysis

**Provide tools and technology for customer data analysis
Multiplatform, multivariate, user-defined inputs and outputs**

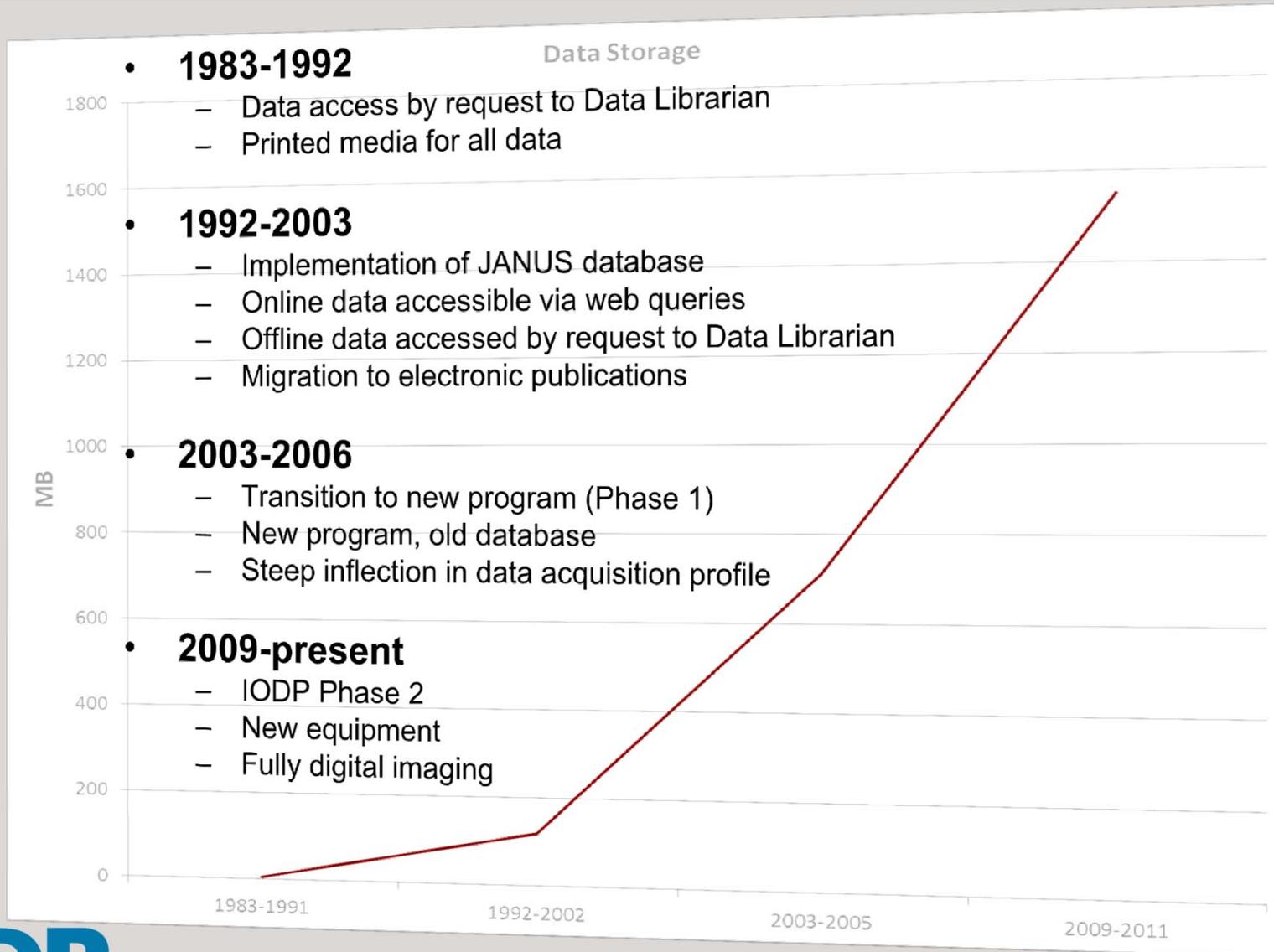
Data archiving

Archive raw and processed data in accessible formats

Data publishing

Provide tools and technology for data discovery and access

Evolution of IODP data management



What was our data dilemma?

Mandate from Stakeholders:

- **1983 - 2003**

- Data collection
- Data archiving
- Data access

**For all data associated with our expeditions
Only subset actually provided online (prime data)**

- **2004 - present**

- All data collected, archived, and accessible
 - Including data from systems we don't manage
 - Real-time data visualization
 - User data entry/editing
 - Remote access to data capture systems
 - User-configurable data discovery tools



Integrated Ocean Drilling Program
United States Implementing Organization

Other considerations

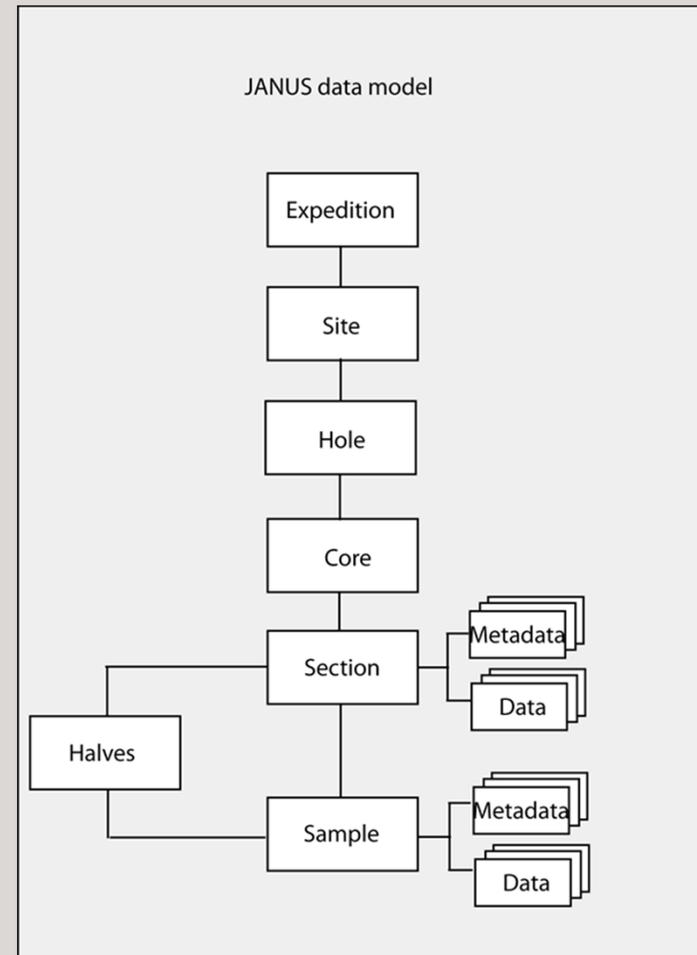


- **Legacy database (JANUS)**
 - Process based
 - Only a small subset of collected data included
 - Tables for each data type
 - Processes constantly changing
 - 1980's database architecture
- **Most complex and subjective data set not included**
 - Earth materials descriptive and interpretive information
- **Variance in scientist core description**
 - New to collaborative effort
 - Differences in experience, culture, education
 - Classic methodology is uniquely individualistic
 - No universally accepted standard, even in basic terminology

Database options (Keep-Revise-Replace)

Keep

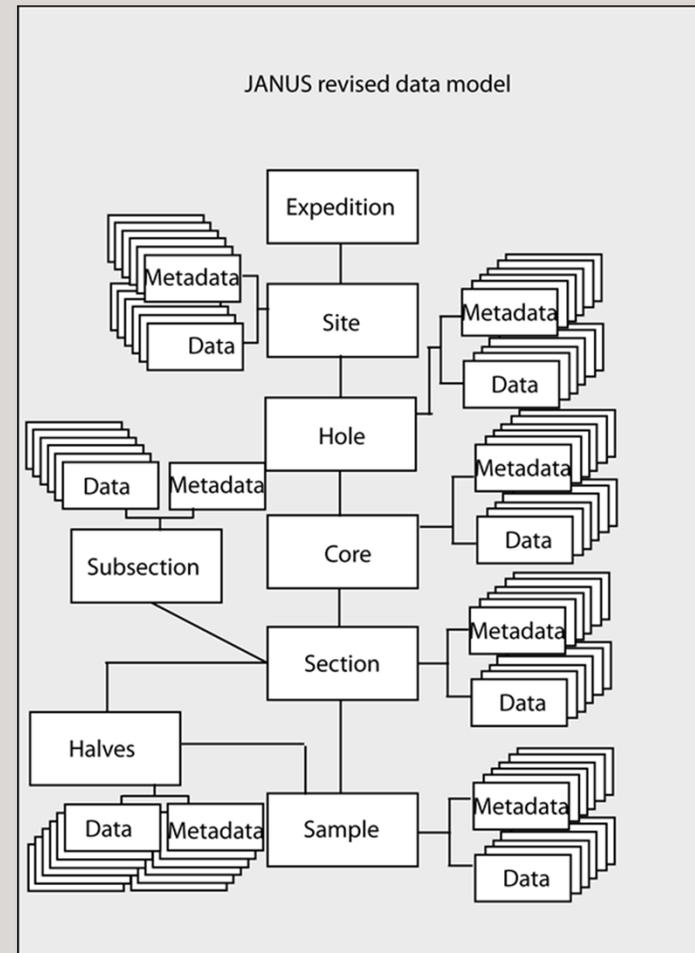
- No new measurements or instruments, no new capabilities
- Loss of some capabilities as new instruments come on line
- Best estimate was <35% of new science data could be accommodated
- Never really seriously considered



Database options (Keep-Revise-Replace)

Revise

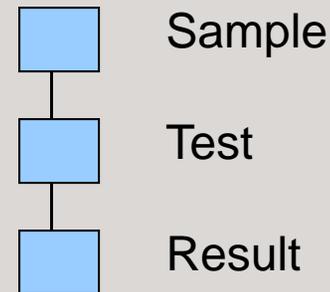
- Add/change virtually every table without upgrading data structure
- Cannot easily provide multiexpedition search capability
- Cannot provide for descriptive information capture
- Cannot provide for digital image management
- CDEX (Japanese implementing organization) attempted to rewrite a JANUS-style data model
 - >70 person-years in development
 - Limited data access, poor performance, insufficient tools for our deliverables



Database options (Keep-Revise-Replace)

Replace

- New, simplified data model
- Utilize state of the art tools and database structure concepts
- Flexibility and expandability
- Estimated 18 person-years for effort based on leverage of commercial LIMS acquisition



What was our solution?

Change database construct

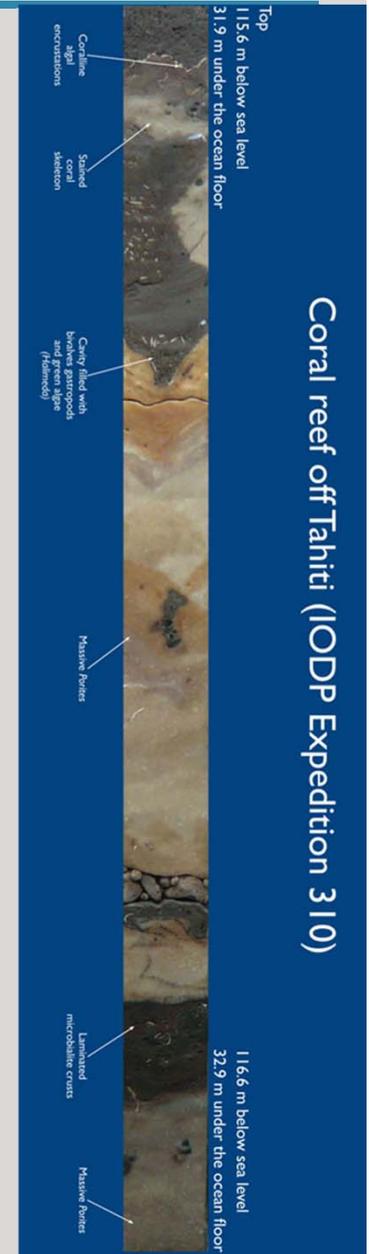
- From process to information based
- Isolate data model from applications
- Consistent, simplified structure
- Allow for incorporation of new tools and/or processes without disrupting data model or data flow

Buy where possible, build only when not

- Market surveys and test groups for commercial products
- If commercial products only provide partial solution - review criteria
- Provide familiar user environment when possible
- User-configurable graphic user interface

Incorporate single most neglected data set

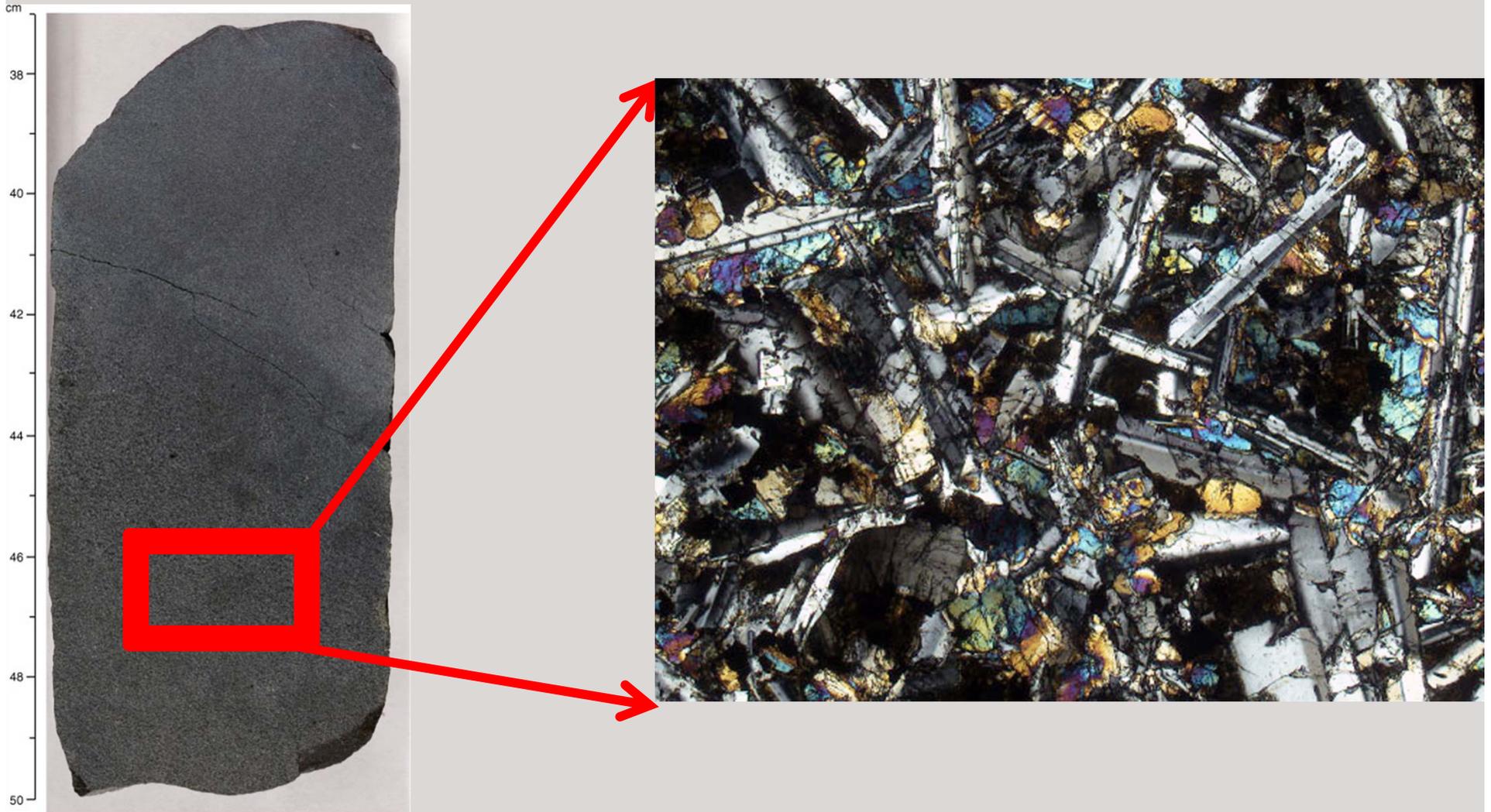
- Earth materials descriptive and interpretive information



Integrated Ocean Drilling Program
United States Implementing Organization

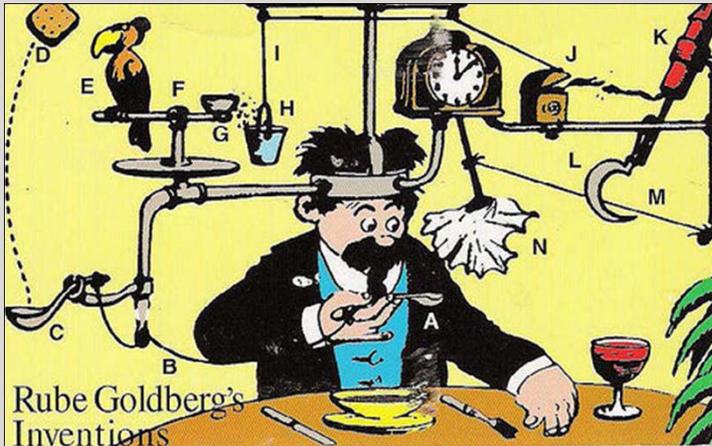
New Database Architecture

The cost of flexibility is complexity - code allowing user simplicity embeds all the complexity of previous developments under the hood.



Simplicity vs. complexity

Complex



Ease to build
Ease to understand



Conservation of matter

$$E = mc^2$$

Simple

Basic

Advanced

Power and utility

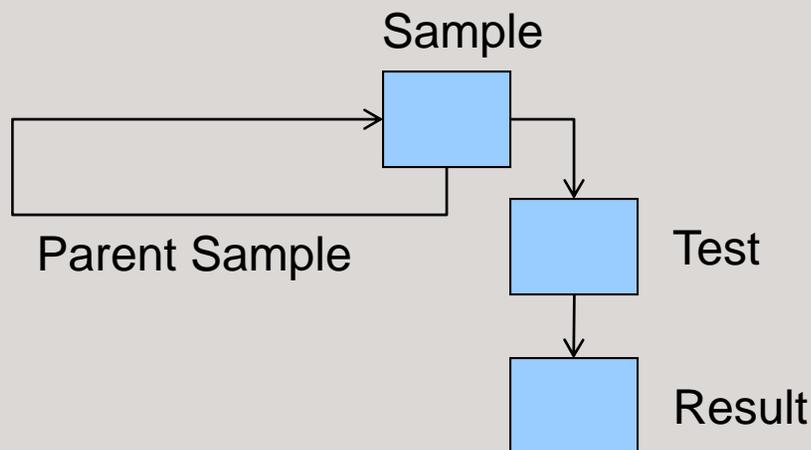


Integrated Ocean Drilling Program
United States Implementing Organization

New Database Architecture

PREMISE #1 – Flexibility - Anything can be a sample

- The hole we drill
 - The core we collect from the hole
 - A piece of the core we collect from the hole
 - An extract of a piece of the core we collect from the hole

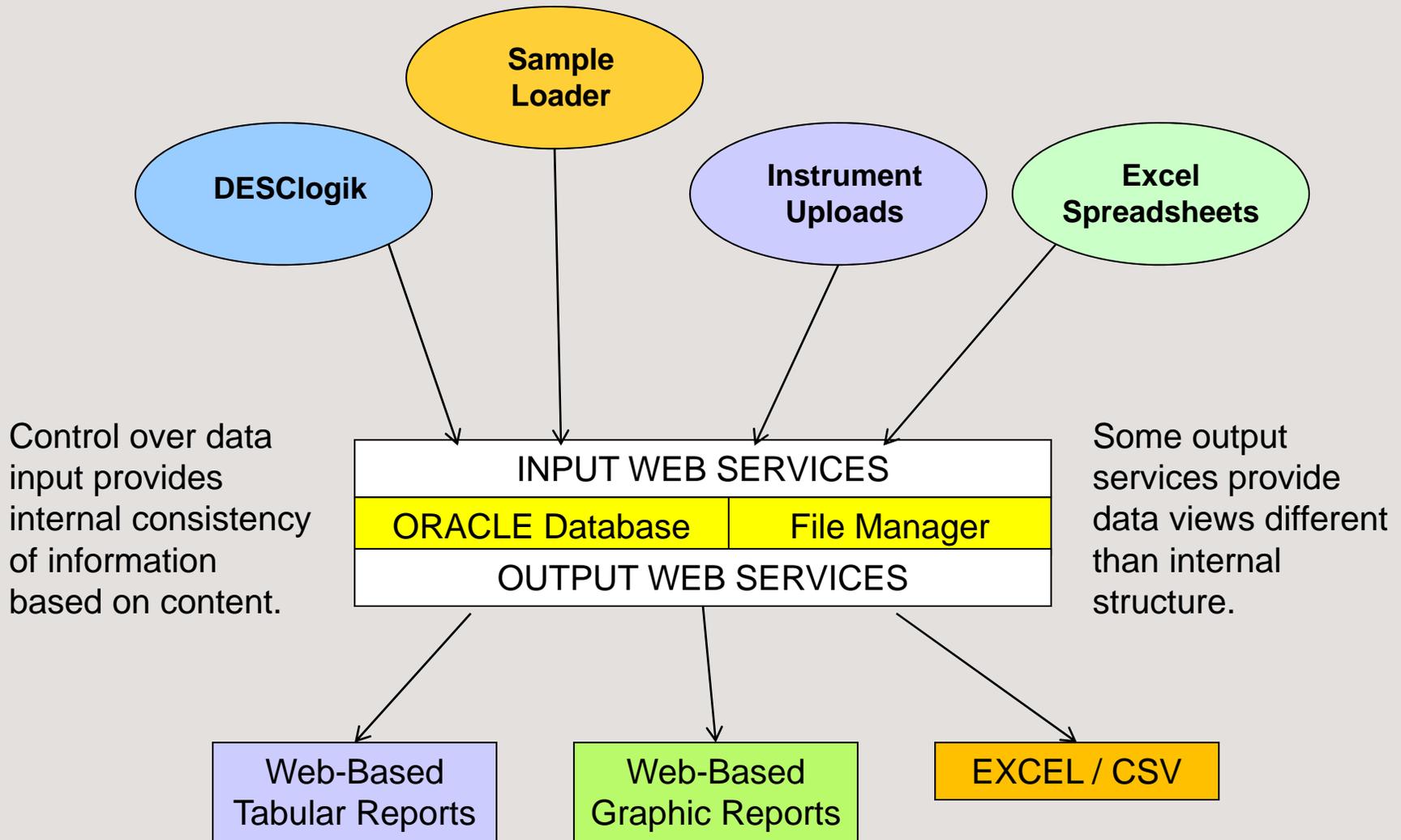


PREMISE #2 - Name – value pairs are used in results to allow many different kinds of information to be managed together:

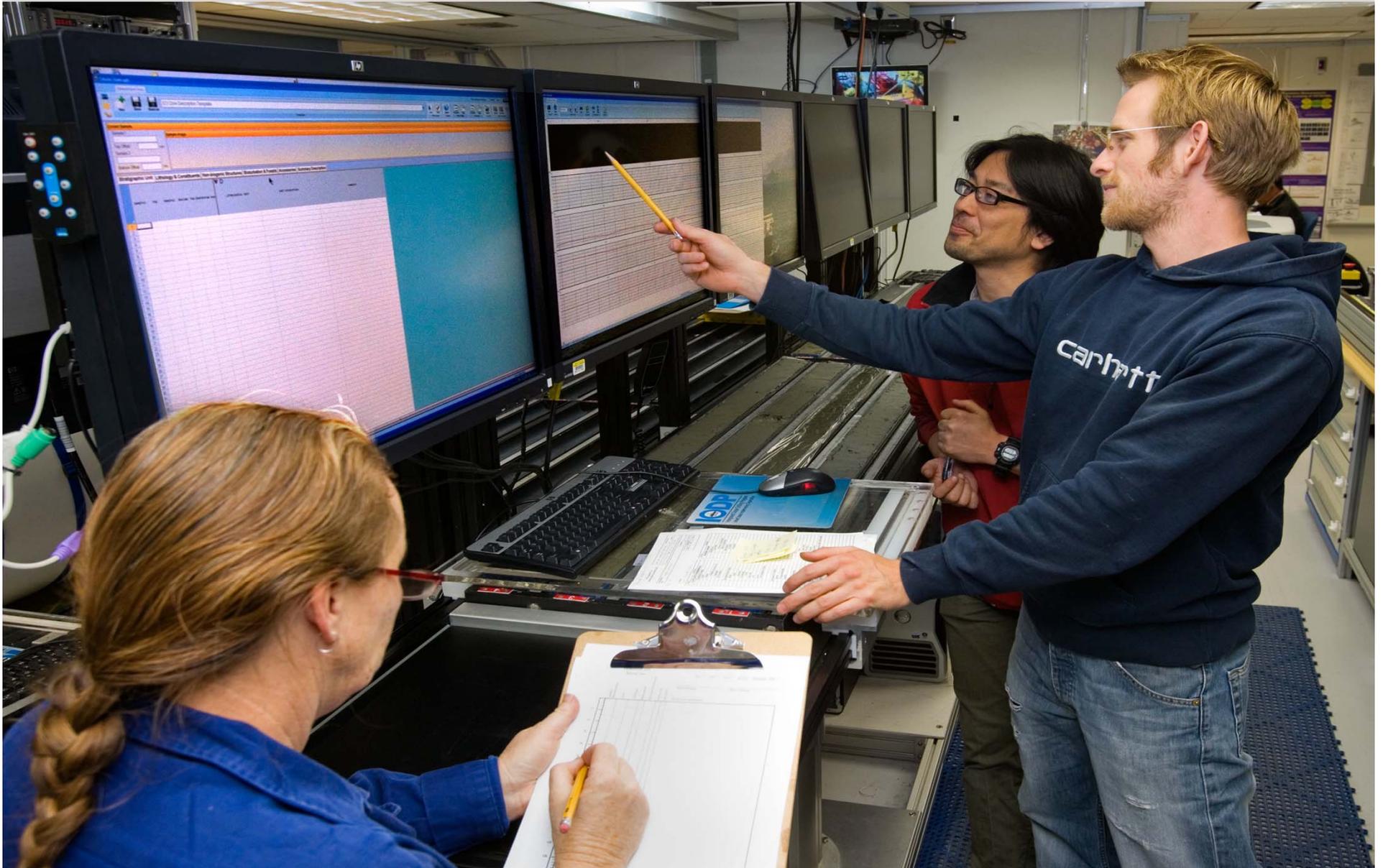
**Example: density 4.5
methane 0.23**

Also allows cross-referencing of same measurements taken by different tests.





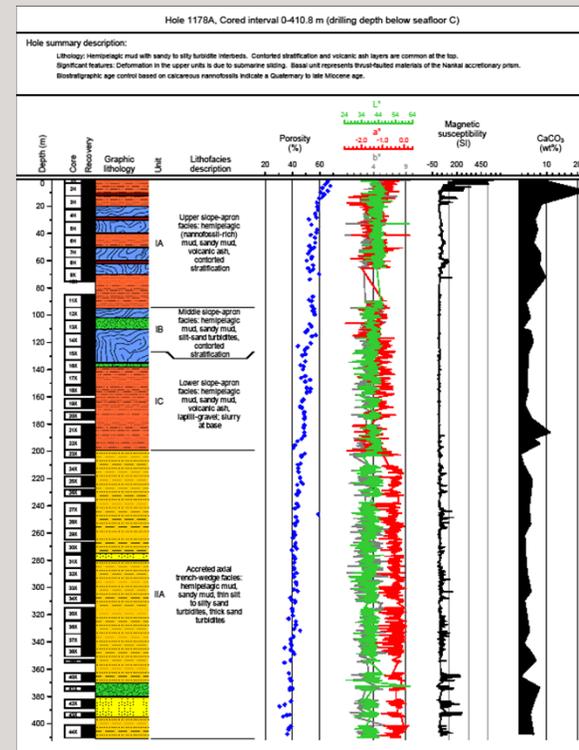
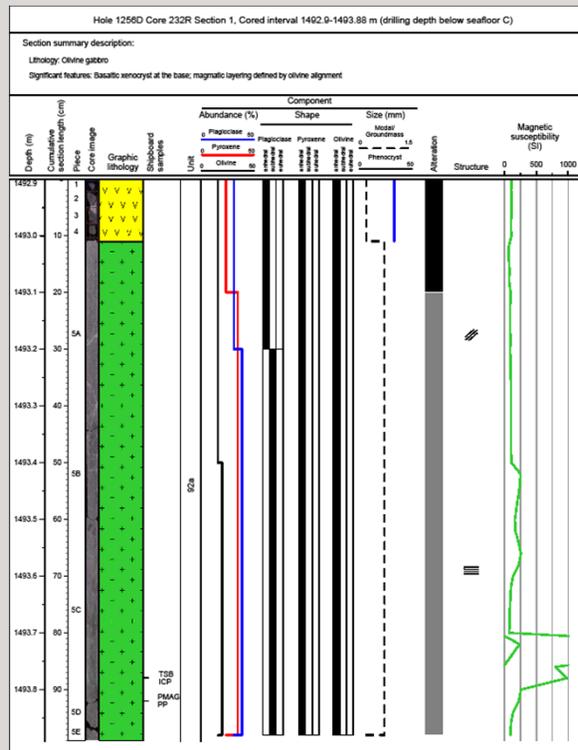
Earth Materials Descriptive and Interpretive Information



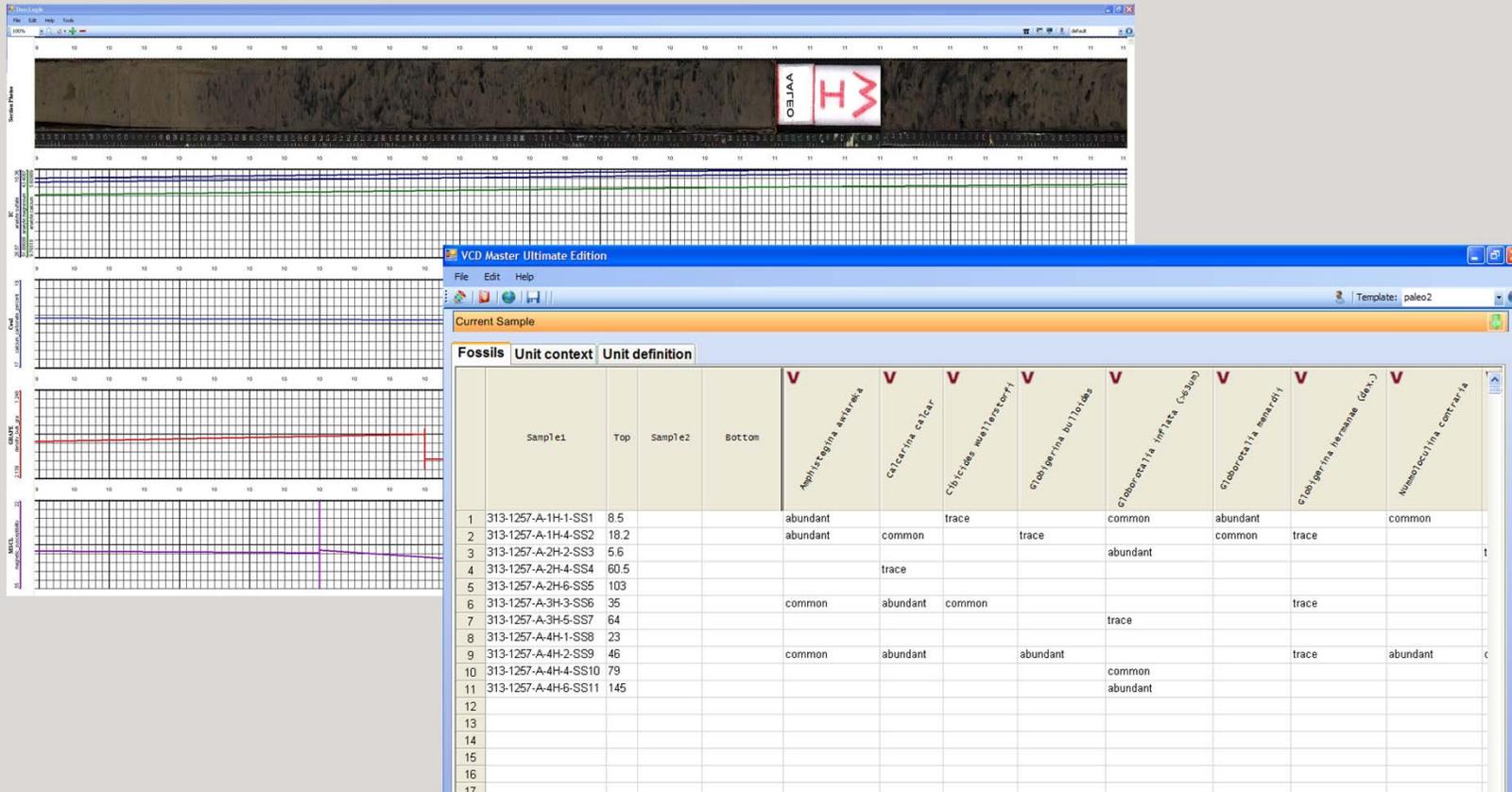
Earth Materials Descriptive and Interpretive Information



Earth Materials Descriptive and Interpretive Information



Earth Materials Descriptive and Interpretive Information



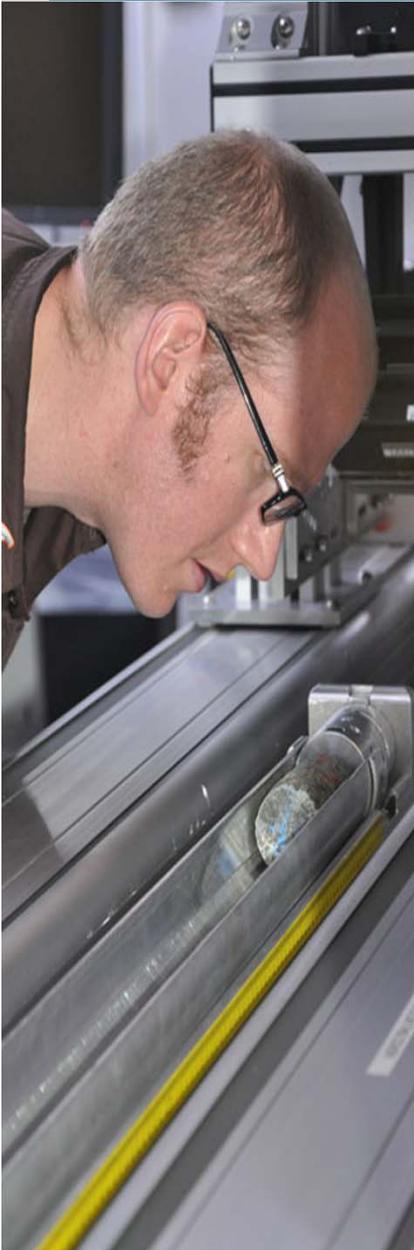
Provide familiar environment

Combine information in visual format with data input

Tools to simplify complex database requirements



What have we learned?



- **Don't undervalue customer expectation management**
- **Don't compromise customer satisfaction**
- **Removing rules increases flexibility, and chaos**
- **Reeducate customers for each use**
- **Include customer education in projections**
- **Avoid back-fitting rules/restrictions**
- **Formatted, simple data retrieval is fundamental**
- **Even if development time is limited, remember getting data out is just as important as getting it in**

What is our vision?



- **Improve customer expectation management**
 - Video tutorials**
 - Workshops**
 - Exposure to new software**
- **Reduce required developer support**
 - Common software architecture**
 - Educate support staff**
 - User data editing tools**
- **Incorporate new data discovery tools**
 - User configurable**

What is our vision?

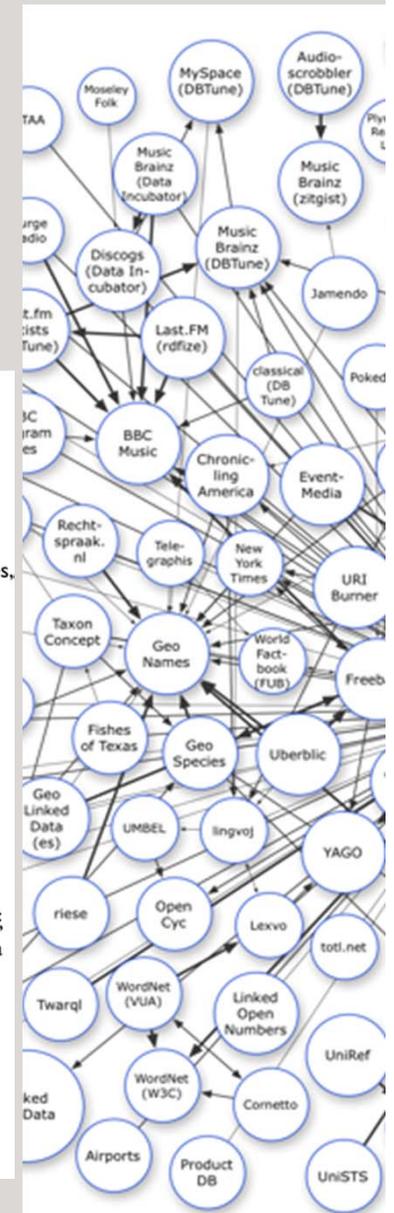
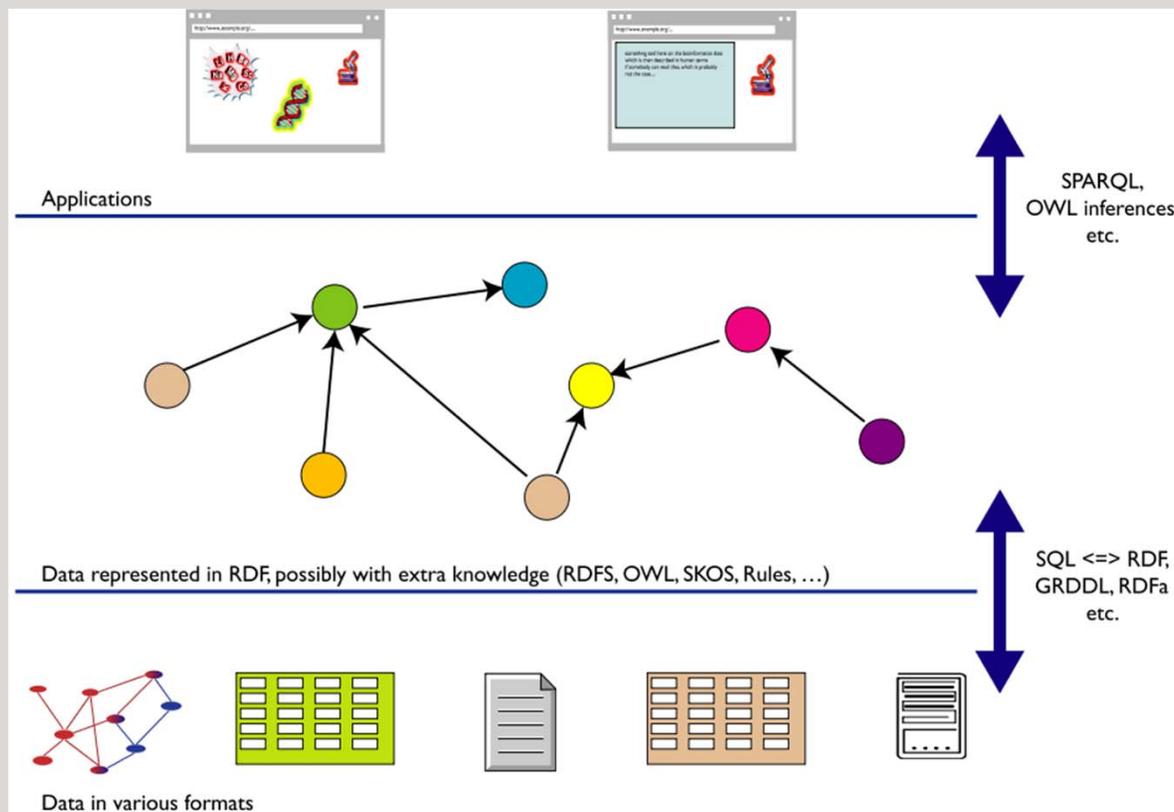
- **Transition to use-based database philosophy**
Change from “how do we capture and present data” to “how do our customers use data”
- **Provide modifications without disrupting data model**
Change core
Change wrapper
Maintain flexibility
- **Migrate all existing data from previous data models**



Integrated Ocean Drilling Program
United States Implementing Organization

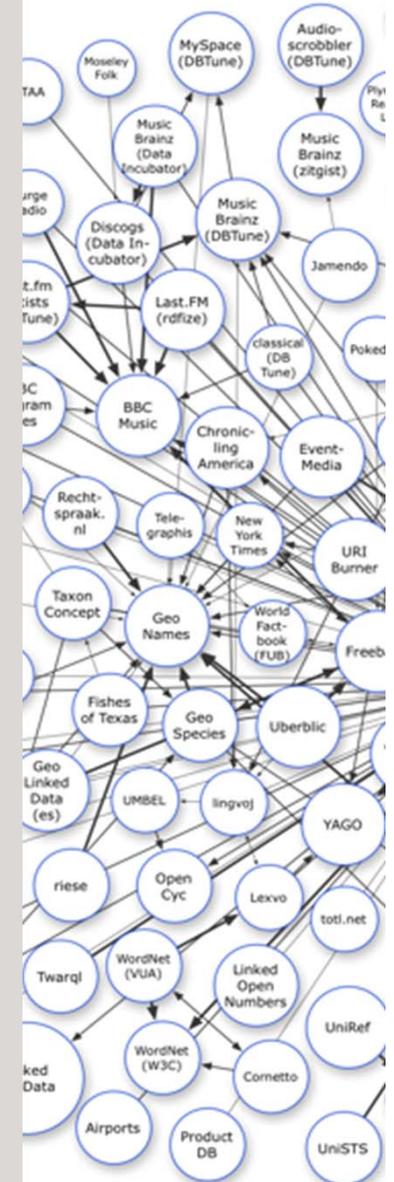
Taking IODP data to Linked Data Space

- Combines
 - IODP-USIO **data** (also ICS, CHRONOS)
 - IODP-IMI funded efforts in **vocabularies**
 - **Linked Data** and Semantic patterns
 - **SKOS (vocab) + RDF (model) + SPARQL (query)**



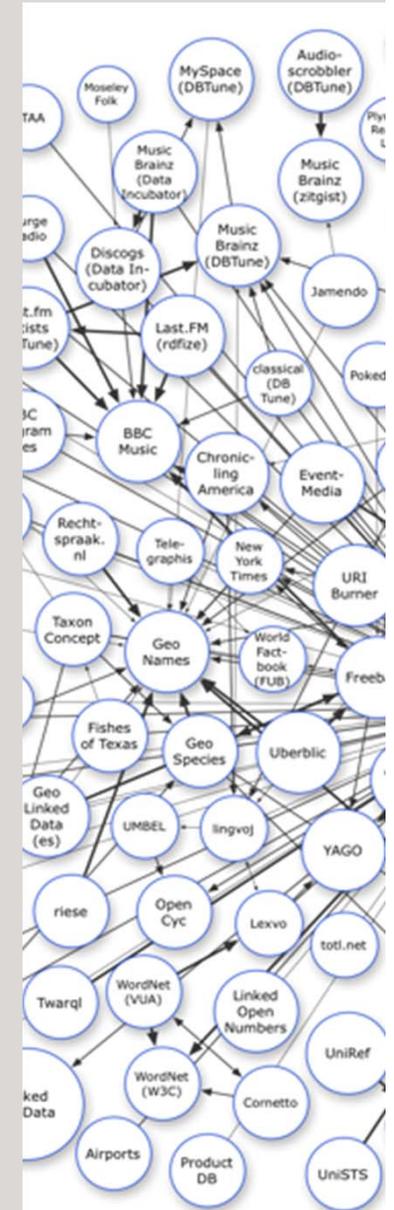
Partners & Status

- Rensselaer Polytechnic Institute (RPI) Tetherless Worlds Constellation (TWC)
 - guidance on vocabulary and linked data patterns
 - provenance and publication linking
 - tools: csv2rdf4lod, eScience (S2S, etc.)
- Google
 - Earth Engine (testing LD architecture as a source for data to Earth Engine)
- Others.. (gplates, Microsoft (Azure Data Services))
- Initial graphs and SPARQL end points services up for testing interfaces and tools at data.oceandrilling.org
- Engage communities to better define how they engage IODP data



Conclusion

- **USIO-IODP data is well positioned**
 - Well structured data
 - Domain science vocabularies used in the system (becoming well structured vocabularies)
 - Community of scientists and users
- **Testing phase now**
 - Working with RPI and others on architecture
 - Evaluating the benefits and effort
 - Engaging community to assess access paths to data



Discussion



–Contact information

Dr. Jay Miller

**Manager, Technical and
Analytical Services**

miller@iodp.tamu.edu

Paul Foster

**Supervisor, Applications
Development**

foster@iodp.tamu.edu

Doug Fils

Data Management

Technical Expert

dfils@oceanleadership.org