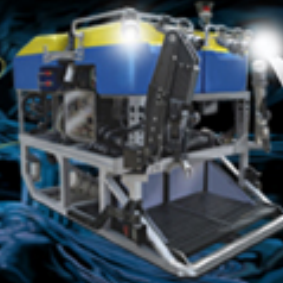
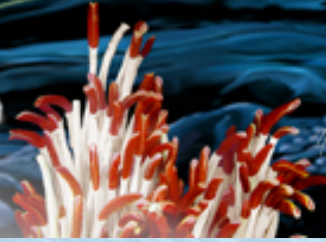
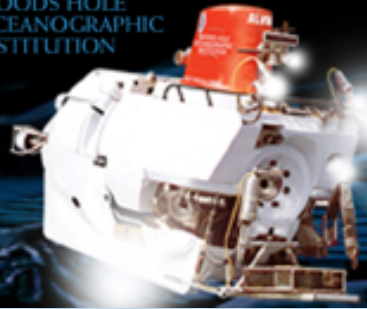




WOODS HOLE  
OCEANOGRAPHIC  
INSTITUTION



# **DESSC Early Career Scientist Program:** *Data Management Overview*

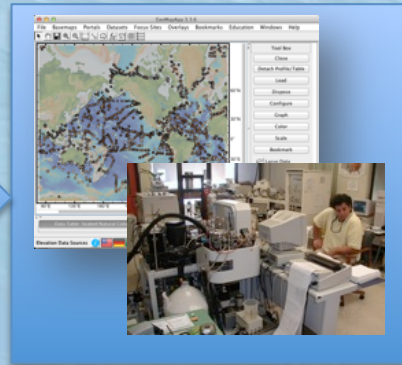
Vicki Ferrini

*Lamont-Doherty Earth Observatory*

# Typical Scientific Workflow



Data Acquisition



Data Processing  
& Interpretation



Publication

# Increasing Emphasis on *Open Data Access*

- Acquisition Costs
- Spatial & Temporal Change
- Scientific Reproducibility
- Federal Data Policy Compliance
- *New Possibilities*
- *Big Data*

Division of Ocean Sciences  
Sample and Data Policy



National Science Foundation

May 2011

News

## White House issues directive supporting public access to publicly funded research

Timothy Vollmer, February 22nd, 2013



Seal of the United States Office of Science and Technology Policy / Public Domain

Today, the White House issued a [Directive](#) supporting public access to publicly-funded research.

John Holdren, Director of the Office of Science and Technology Policy, "has directed Federal agencies with more than \$100M in R&D expenditures to develop plans to make the published results of federally funded research freely available to the public within one year of publication and requiring researchers to better account for and manage the digital data resulting from federally funded scientific research."

Each agency covered by the [Directive](#) (54 KB PDF) must "Ensure that the public can read, download, and analyze in digital form final peer reviewed manuscripts or final published documents within a timeframe that is

appropriate for each type of research conducted or sponsored by the agency."

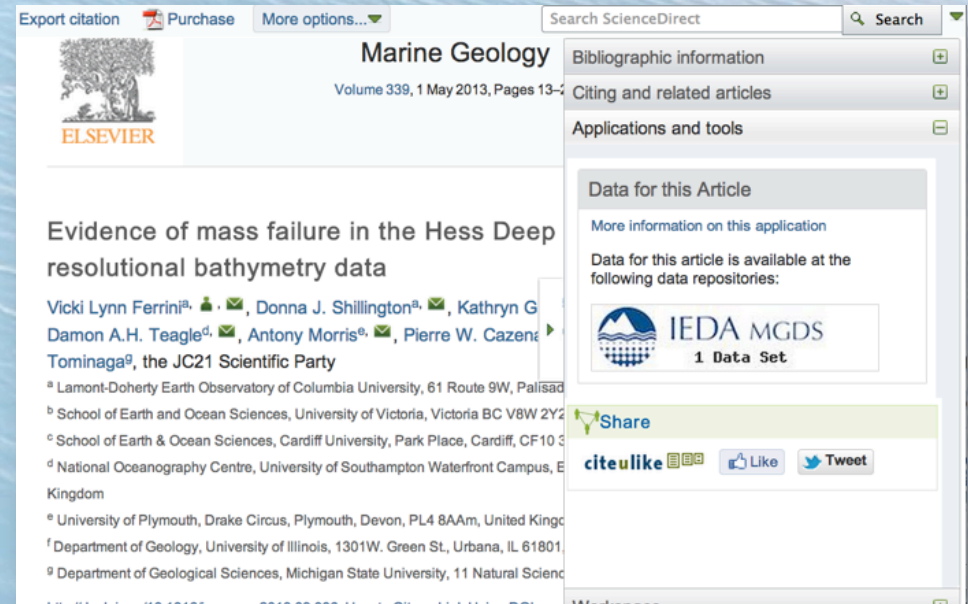
A screenshot of the National Science Foundation website homepage. The header features the NSF logo and the text 'National Science Foundation WHERE DISCOVERIES BEGIN'. Below the header is a navigation menu with links for HOME, FUNDING, AWARDS, DISCOVERIES, NEWS, PUBLICATIONS, STATISTICS, ABOUT, and FastLane. A search bar is located in the top right corner. The main content area features a large banner with the text 'Cloud Computing Opens New Possibilities' and a 'FEATURES' section with numbered links 1, 2, 3, 4.

# Why now?

- Exponentially increasing data volumes
- Rapidly expanding cyberinfrastructure capabilities to mine and analyze data
- Need for cross-domain data access & integration
- New paradigms in publishing
- Open access requirements from funders

# What's in it for you?

- Scientific Integrity & Reproducibility
- Opportunity
- Attribution
- Increase the impact of your research
- Preserve data for your own future use
- Compliance with Data Policies
- Education & Outreach



“The coolest thing to do with your data will be thought of by someone else.”

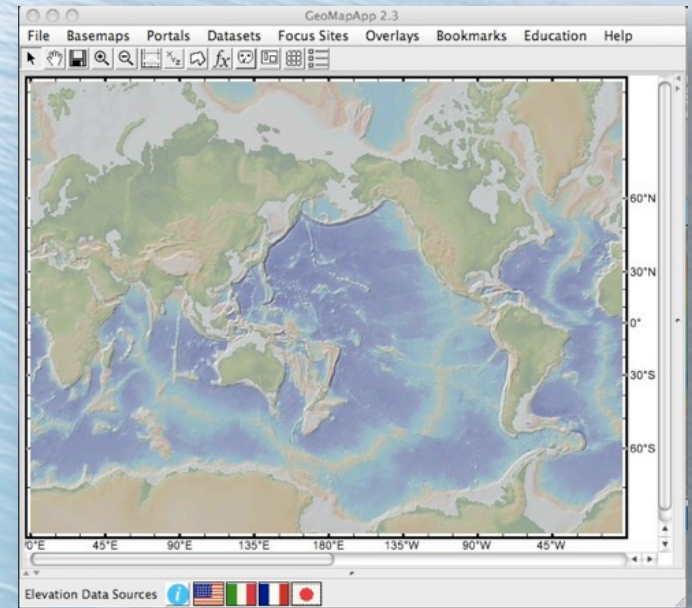
*Rufus Pollock*


*Cambridge University and Open Knowledge Foundation*



# Plan

- Concept/Proposal Development
  - *Are Existing Data Available?*
- Data Acquisition Plan
  - Sensor Calibration
  - Survey Plan
  - Data Analysis + reduction
  - Data Documentation
- Data Management Plan (DMP)
  - *How will you preserve & document your data?*



 **INTEGRATED EARTH DATA APPLICATIONS**  
seddata.org

### Data Management Plan

**Primary Investigator:** John Morton  
**Institution:** Lamont Doherty Earth Observatory of Columbia University  
**Project:** Reactivation of the Passive Margin of Eastern Laurentia  
**NSF Division:** OCE **Solicitation Info:** Marine Geology and Geophysics **Submission Date:** 01/16/2013

**Overview:** Our project will use active source seismology on the Marcus G. Langseth to image the oceanic crust on the continental shelf of the Eastern U.S. after the Dec. 21, 2012 earthquake.

**Data description:** The proposed research will result in several new seismic transects along and across the new active margin.

**Data analysis summary:** CMP stacking and migration will be performed using the open source seismic utilities package Seismic Unix. Gravity data will be processed using the open source R2R\_Gravity data processing tools. Multibeam bathymetry will be processed using MBSsystem.

**Includes field work?** Yes

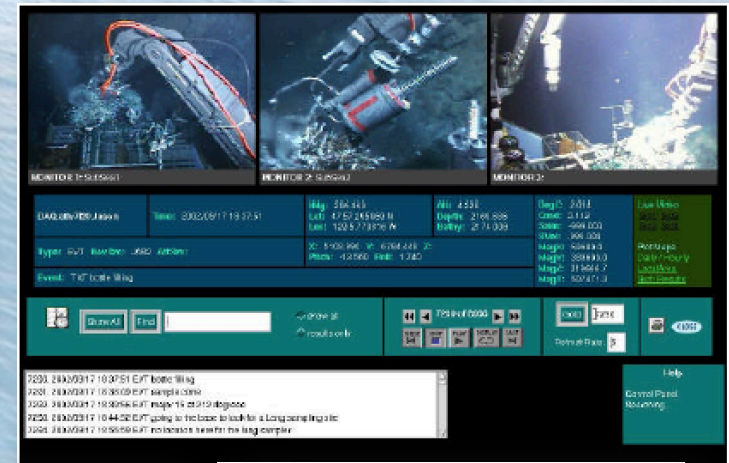
**Description of field work:** Active source seismology, multibeam bathymetry, and gravimetry (BGM-3) data will be collected.

**Expected data product #1**  
**Data type:** Observational, Analytical  
**Responsible investigator:** John Morton  
**Product description:** .segv files from seismic transits.  
**Intended repository:** IRIS  
**Timeline for data release:** Immediate Release

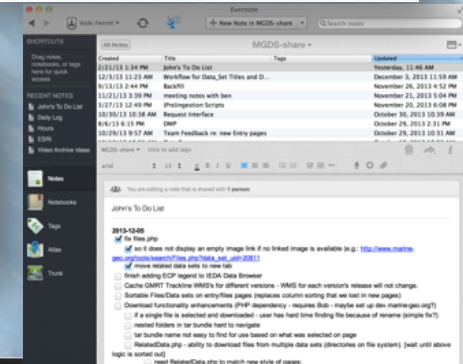
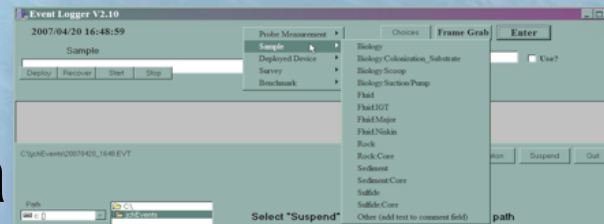
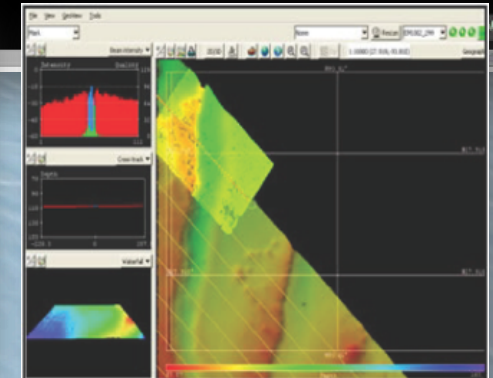
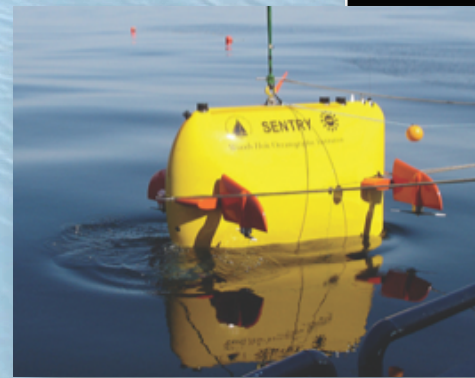
**Expected data product #2**  
**Data type:** Observational  
**Responsible investigator:** John Morton  
**Product description:** Processed free-air anomaly data in MGD77-T format  
**Intended repository:** NGDC  
**Timeline for data release:** Immediate Release

**Expected data product #3**  
**Data type:** Observational  
**Responsible investigator:** Vicki L. Ferrini  
**Product description:** Multibeam bathymetry data  
**Intended repository:** MGDS  
**Timeline for data release:** Immediate Release

# Collect & Assure



- Operational Limitations
- Technical Limitations
- Scientific Standards
- Sensor Performance
- Quality/Coverage Assessment
- *Contemporaneous* data documentation
- Opportunistic data acquisition

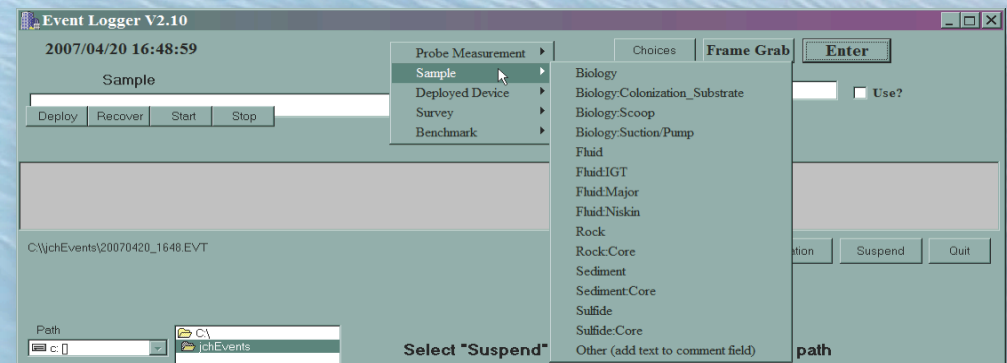




# Document & Preserve

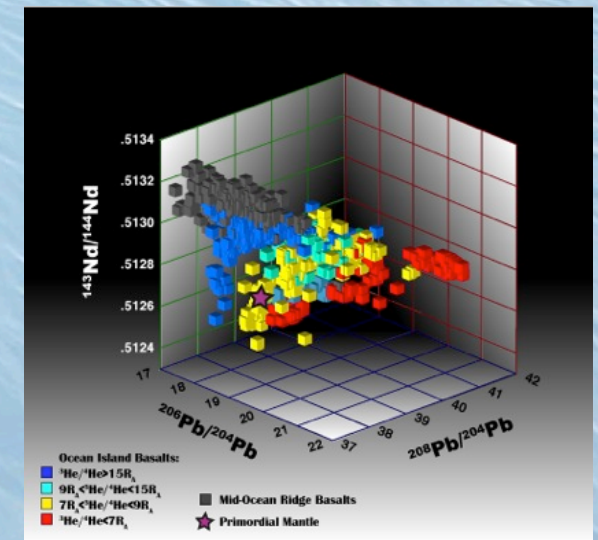
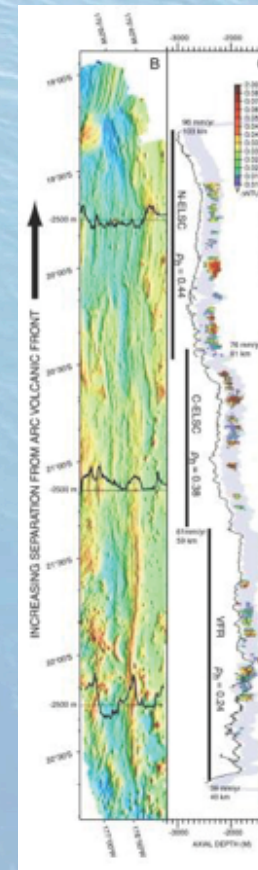
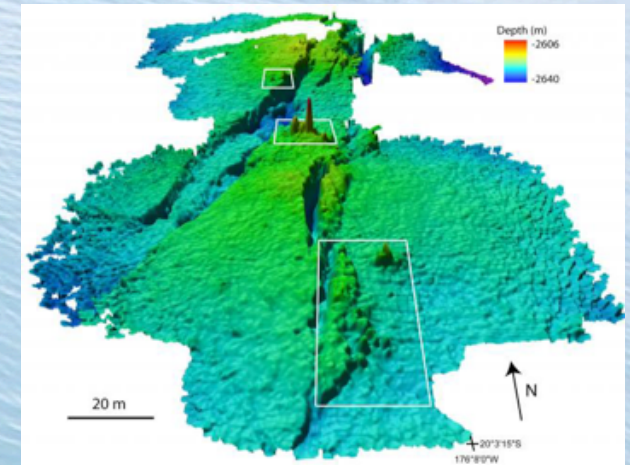
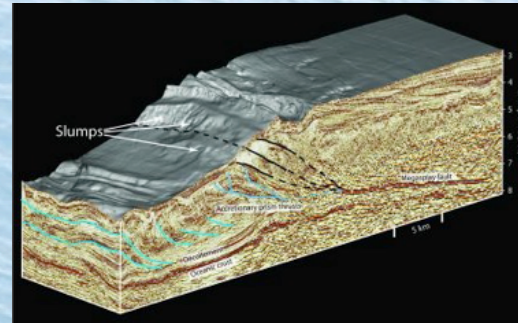


- Cruise report
- **Raw sensor data** ✓
- Science party instrumentation ?
- Sample metadata
- Physical samples



# Analyze

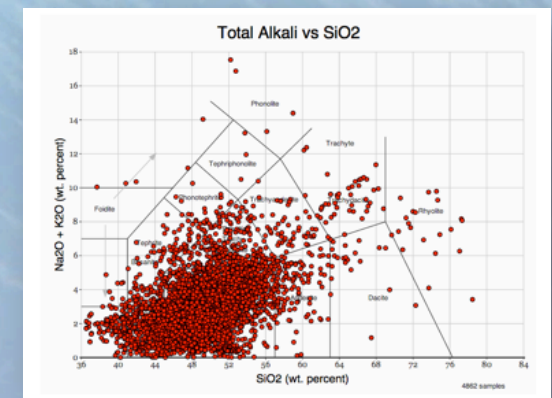
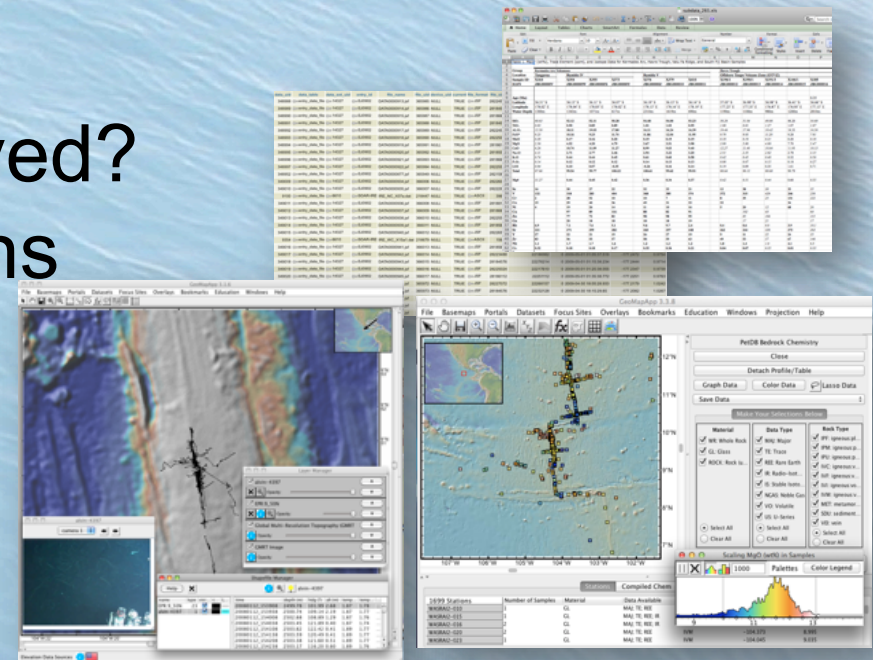
- Process Samples
- Reduce Data
- Document
  - Assumptions
  - Technical Limitations
  - Versioning
  - Protocols
  - Scientific Standards
  - Quality



# Document & Preserve

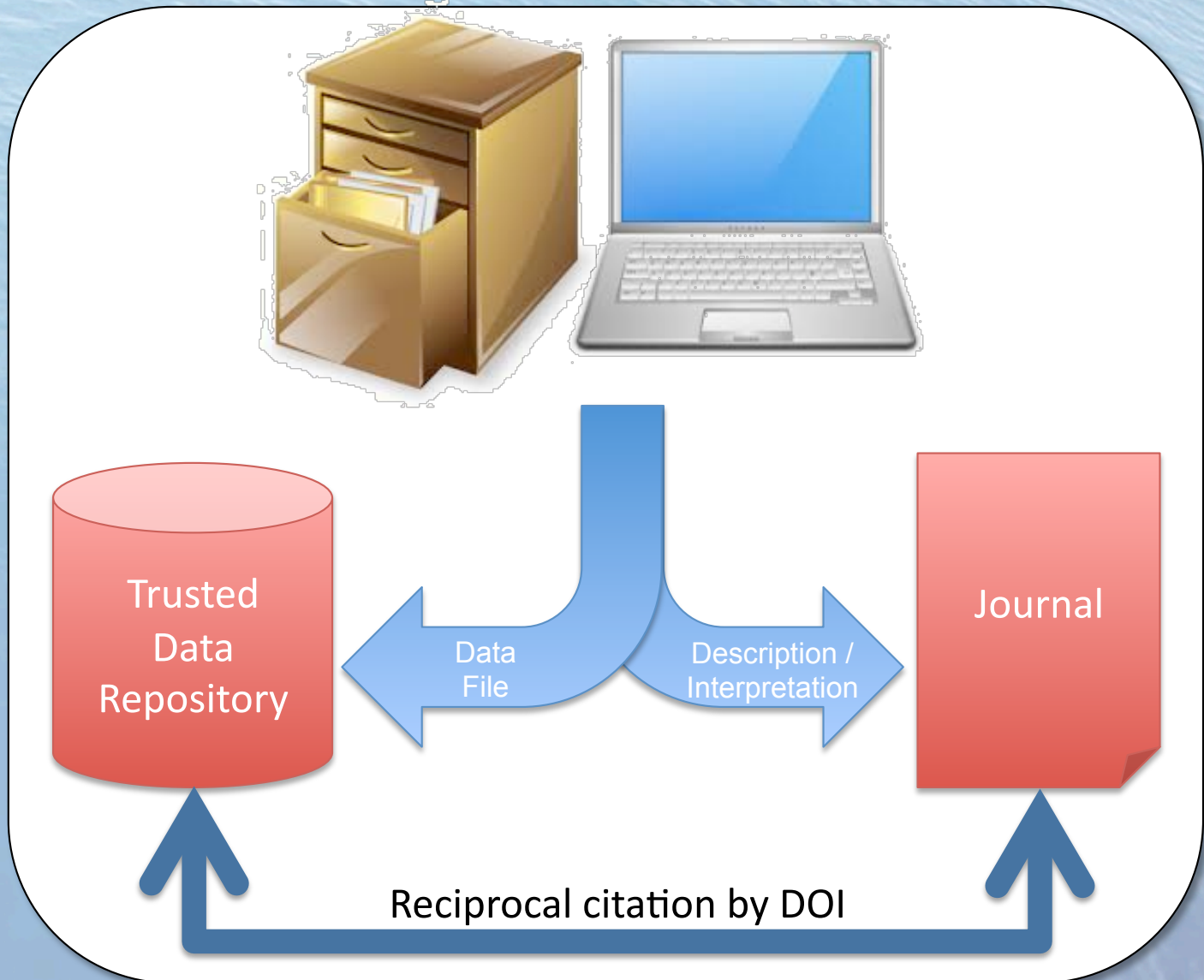


- Which data should be preserved?
  - Data supporting publications
  - Processed data of value
  - Results of lab analysis
- Where should it be curated?
- Documentation
  - What does a new user need to know?
  - How were products generated?
  - What are caveats of data?



**Integrate &  
Share**

**“Best Practice”**

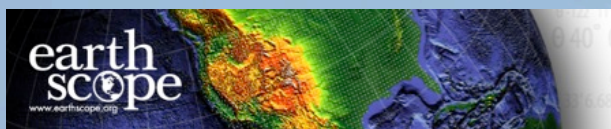
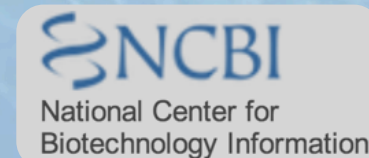






# Which Repository?

- Know data policies
- Seek domain-specific repositories
- System Features
  - Data Usage Reports
  - Data Publication
  - User support
  - Usability
  - Interoperability





# Discussion