Multi-Institution Testbed for Scalable Digital Archiving

NSF CISE/Library of Congress DIGARCH Award

Stephen Miller
Scripps Institution of Oceanography

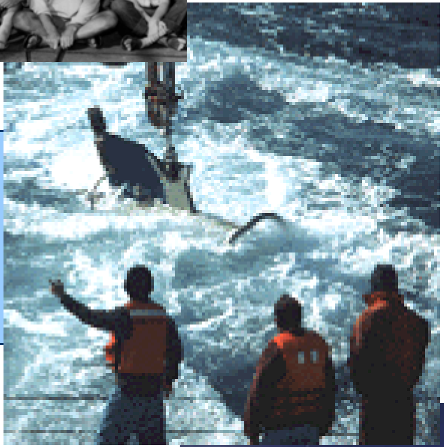
Bob Detrick
Woods Hole Oceanographic Institution

John Helly
San Diego Supercomputer Center

dg.o 2005 THE NATIONAL CONFERENCE ON DIGITAL GOVERNMENT RESEARCH

Atlanta 2005-05-17

1. Community Goals



2. Barriers to advances

3. Cyber-capabilities

1. Community Goals

Broad support

Across disciplines

And institutions

Research

And education

Guarantee long term preservation

Gulf of California 1939 Expedition, R/V E W Scripps

Need more than data storage

Need metadata

Enable re-use

Also need infrastructure

Networked community tools, archives, understanding

Why re-use data?

New ship time expensive ($22K/day)

Use archives for:

1. Regional synthesis projects

3. Support other disciplines

3. Monitor environmental changes through time

Before and after

Earthquakes, slumps, seeps

Volcanoes …

SIOExplorer

April 16, 2005

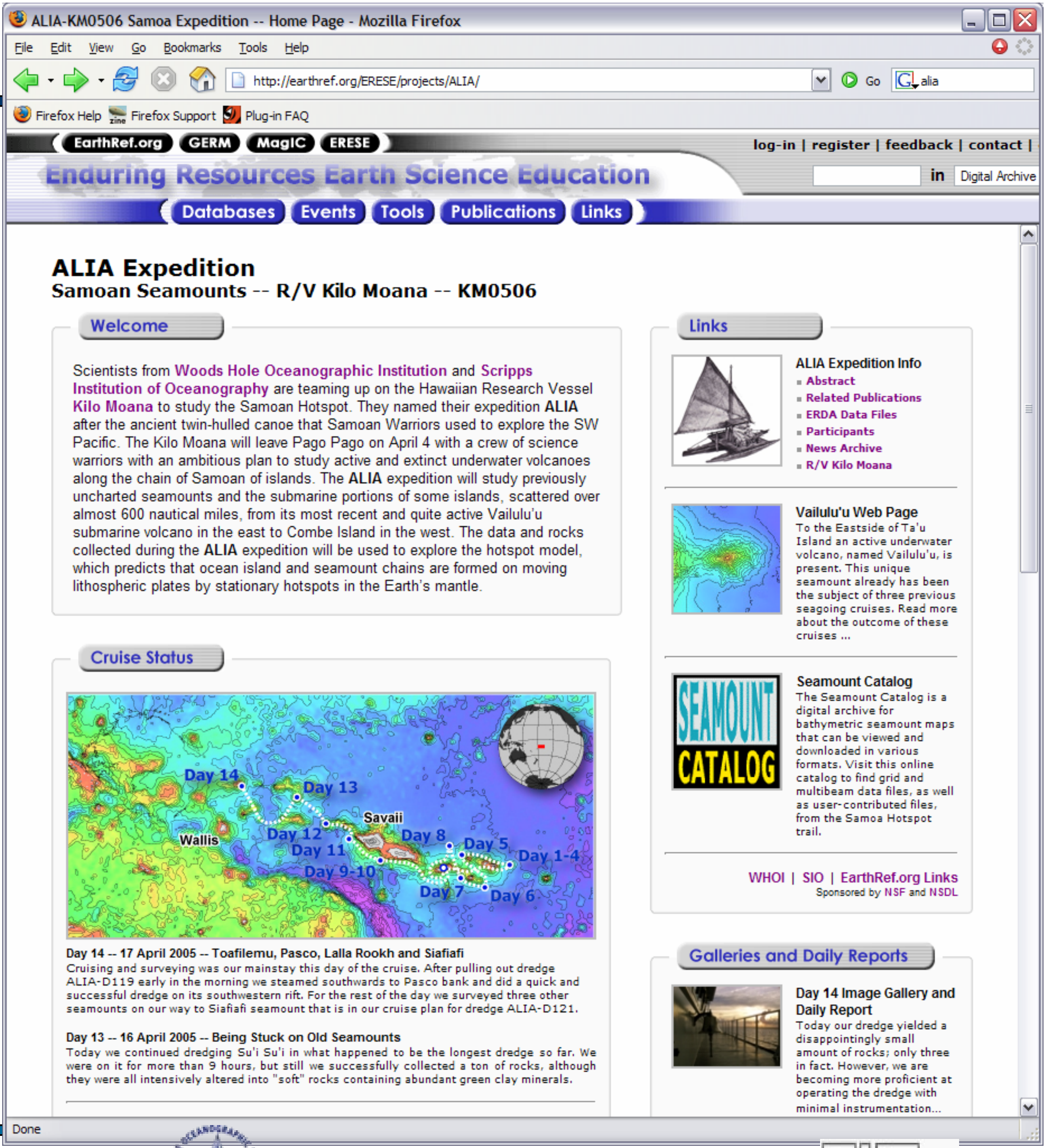
New Volcanic Cone in the Vailulu'u Crater

With a minimum rate of **eight inches per day**, a new cone has been growing inside the crater of Vailulu'u seamount since the last depth soundings by the US Coastguard vessel Polar Sea in April 2001.

Our survey using the SIMRAD 120 system of the Kilo Moana displays a new volcanic summit at 708 m depth.

This volcano was named Nafanua, after the Samoan Goddess of War.

Hubert Staudigel, Stan Hart, John Helly, Anthony Koppers, Jasper Kontor


ALIA-KM0506 Samoa Expedition -- Home Page - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://earthref.org/ERESE/projects/ALIA/
Go
alia
Firefox Help
Firefox Support
Plug-in FAQ
EarthRef.org
GERM
MagIC
ERESE
log-in | register | feedback | contact |
Enduring Resources Earth Science Education
in
Digital Archive
Databases
Events
Tools
Publications
Links
Day 14
Day 13
Savaii
Wallis
Day 12
Day 11
Day 9-10
Day 8
Day 7
Day 5
Day 6
Day 1-4
Done

ALIA Expedition

Samoan Seamounts -- R/V Kilo Moana -- KM0506

Welcome

Scientists from Woods Hole Oceanographic Institution and Scripps Institution of Oceanography are teaming up on the Hawaiian Research Vessel Kilo Moana to study the Samoan Hotspot. They named their expedition **ALIA** after the ancient twin-hulled canoe that Samoan Warriors used to explore the SW Pacific. The Kilo Moana will leave Pago Pago on April 4 with a crew of science warriors with an ambitious plan to study active and extinct underwater volcanoes along the chain of Samoan of islands. The **ALIA** expedition will study previously uncharted seamounts and the submarine portions of some islands, scattered over almost 600 nautical miles, from its most recent and quite active Vailulu'u submarine volcano in the east to Combe Island in the west. The data and rocks collected during the **ALIA** expedition will be used to explore the hotspot model, which predicts that ocean island and seamount chains are formed on moving lithospheric plates by stationary hotspots in the Earth's mantle.

Cruise Status

Day 14 -- 17 April 2005 -- Toafilemu, Pasco, Lalla Rookh and Siafiafi

Cruising and surveying was our mainstay this day of the cruise. After pulling out dredge ALIA-D119 early in the morning we steamed southwards to Pasco bank and did a quick and successful dredge on its southwestern rift. For the rest of the day we surveyed three other seamounts on our way to Siafiafi seamount that is in our cruise plan for dredge ALIA-D121.

Day 13 -- 16 April 2005 -- Being Stuck on Old Seamounts

Today we continued dredging Su'i Su'i in what happened to be the longest dredge so far. We were on it for more than 9 hours, but still we successfully collected a ton of rocks, although they were all intensively altered into "soft" rocks containing abundant green clay minerals.

Links

ALIA Expedition Info

- Abstract
- Related Publications
- ERDA Data Files
- Participants
- News Archive
- R/V Kilo Moana

Vailulu'u Web Page

To the Eastside of Ta'u Island an active underwater volcano, named Vailulu'u, is present. This unique seamount already has been the subject of three previous seagoing cruises. Read more about the outcome of these cruises ...

Seamount Catalog

The Seamount Catalog is a digital archive for bathymetric seamount maps that can be viewed and downloaded in various formats. Visit this online catalog to find grid and multibeam data files, as well as user-contributed files, from the Samoa Hotspot trail.

WHOI | SIO | EarthRef.org Links

Sponsored by NSF and NSDL

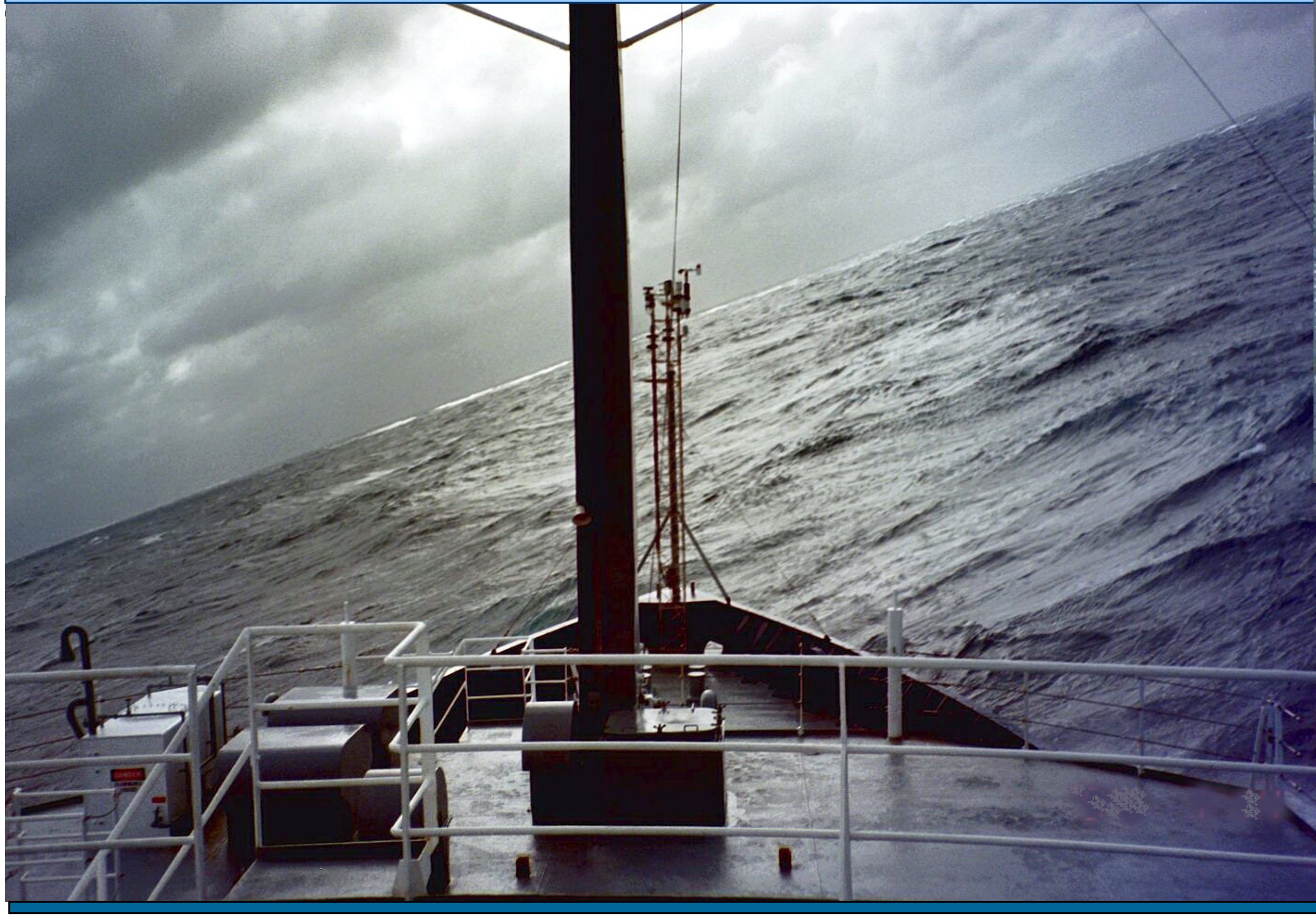
Galleries and Daily Reports

Day 14 Image Gallery and Daily Report

Today our dredge yielded a disappointingly small amount of rocks; only three in fact. However, we are becoming more proficient at operating the dredge with minimal instrumentation...

SDSC

2. Barriers to advances



Data from a firehose

Can we keep up?

Shipboard data rates – **yes**

Satellite links – **maybe**
depends on heading

Metadata – **yes, but**
not widely implemented

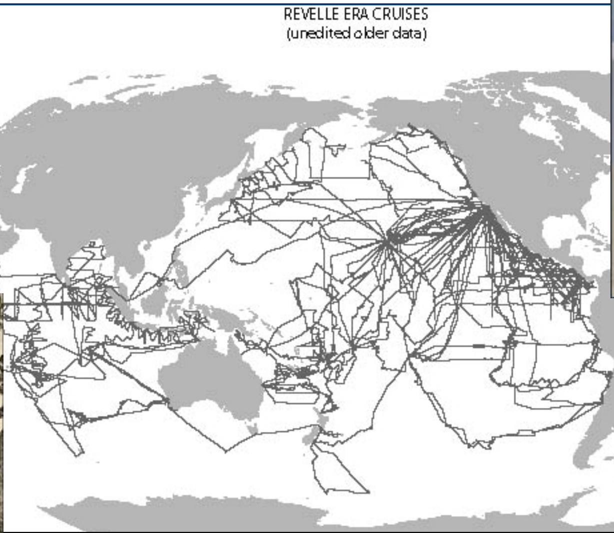
Preservation – **maybe**

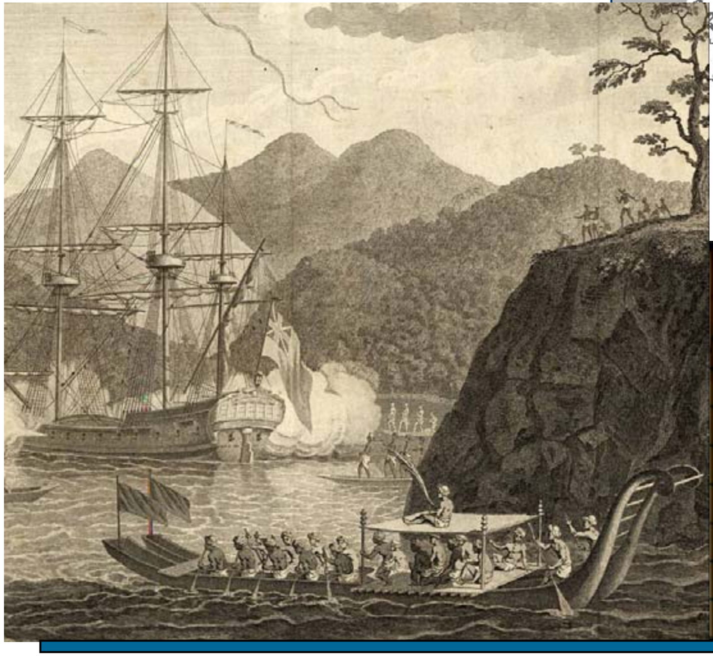
Tiffany Houghton, SDSC, on R/V Sproul

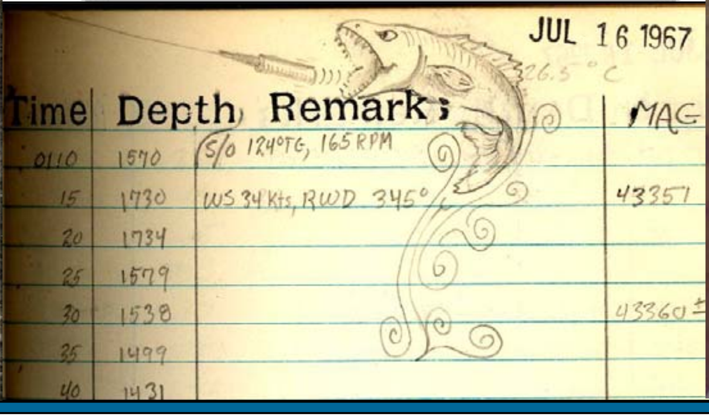
Community usage –**help needed from Cyberinfrastructure**

We can archive from paper documents

- Track plots
- Cruise reports
- Handwritten and printed data

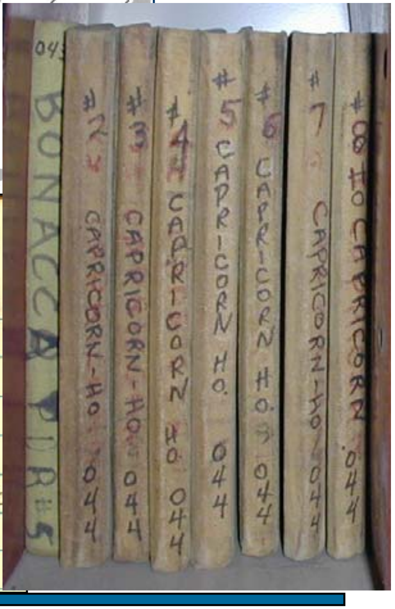

REVELLE ERA CRUISES
(unedited older data)




JUL 16 1967

26.5 °C

Time	Depth	Remark;	MAG
0110	1570	S/O 124°TG, 165 RPM	
15	1730	WS 34 Kts, RWD 345°	43351
20	1734		
25	1579		
30	1538		43360±
35	1499		
40	1431		



But digital preservation is risky business

Endangered Species

9-track tapes →

Exabytes fail

Even CDs fail

RAIDS fail

SIO GEOLOGICAL DATA CENTER
REEL ID GDØ199
LOCATION T04A02
BASF ENDURA Tested for use at 800 BPI through 6250 BPI
BASF
Scripps Institution of Oceanography
R/V Melville
Shipboard Computer Group
SIC:
Paul Robbins
SCG:
Geoff Davis
SeaBeam 2000 Multibeam Data
SCG ID:
VANC32MV
19jun2004 to 20jun2004

“Shoe-box” archiving not to be trusted

Solution: Active Archiving

“Don’t trust any media, person or process”

Actively monitor status

Migrate to new storage media

Mirror on multiple systems daily

Backup to independent sites

Technology makes this possible, just need to do it

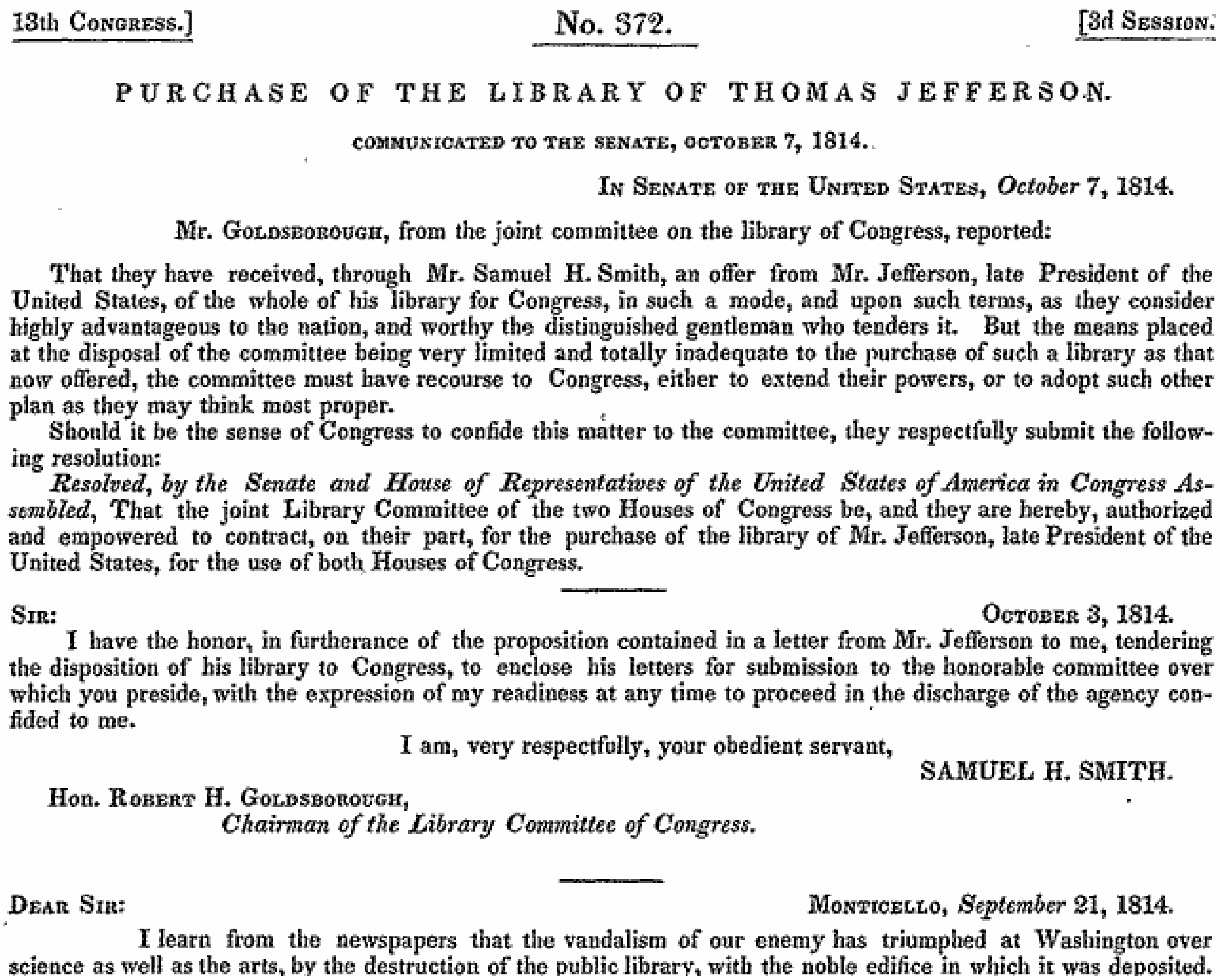
SIOExplorer

Example of early backup

Capital burned August 19, 1814

Library of Congress offsite backup

Thomas Jefferson's Library



13th Congress.] No. 372. [3d Session.

PURCHASE OF THE LIBRARY OF THOMAS JEFFERSON.

COMMUNICATED TO THE SENATE, OCTOBER 7, 1814.

In Senate of the United States, *October* 7, 1814.

Mr. Goldsborough, from the joint committee on the library of Congress, reported:

That they have received, through Mr. Samuel H. Smith, an offer from Mr. Jefferson, late President of the United States, of the whole of his library for Congress, in such a mode, and upon such terms, as they consider highly advantageous to the nation, and worthy the distinguished gentleman who tenders it. But the means placed at the disposal of the committee being very limited and totally inadequate to the purchase of such a library as that now offered, the committee must have recourse to Congress, either to extend their powers, or to adopt such other plan as they may think most proper.

Should it be the sense of Congress to confide this matter to the committee, they respectfully submit the following resolution:

Resolved, by the Senate and House of Representatives of the United States of America in Congress Assembled, That the joint Library Committee of the two Houses of Congress be, and they are hereby, authorized and empowered to contract, on their part, for the purchase of the library of Mr. Jefferson, late President of the United States, for the use of both Houses of Congress.

Sir: October 3, 1814.

I have the honor, in furtherance of the proposition contained in a letter from Mr. Jefferson to me, tendering the disposition of his library to Congress, to enclose his letters for submission to the honorable committee over which you preside, with the expression of my readiness at any time to proceed in the discharge of the agency confided to me.

I am, very respectfully, your obedient servant,
SAMUEL H. SMITH.

Hon. Robert H. Goldsborough,
Chairman of the Library Committee of Congress.

Dear Sir: Monticello, *September* 21, 1814.

I learn from the newspapers that the vandalism of our enemy has triumphed at Washington over science as well as the arts, by the destruction of the public library, with the noble edifice in which it was deposited.

3. Emerging Cyber-capabilities

SIOExplorer digital library

Design for scalability
Automate harvesting
Collection Builder’s Toolkit for other projects

Crossing institutional boundaries

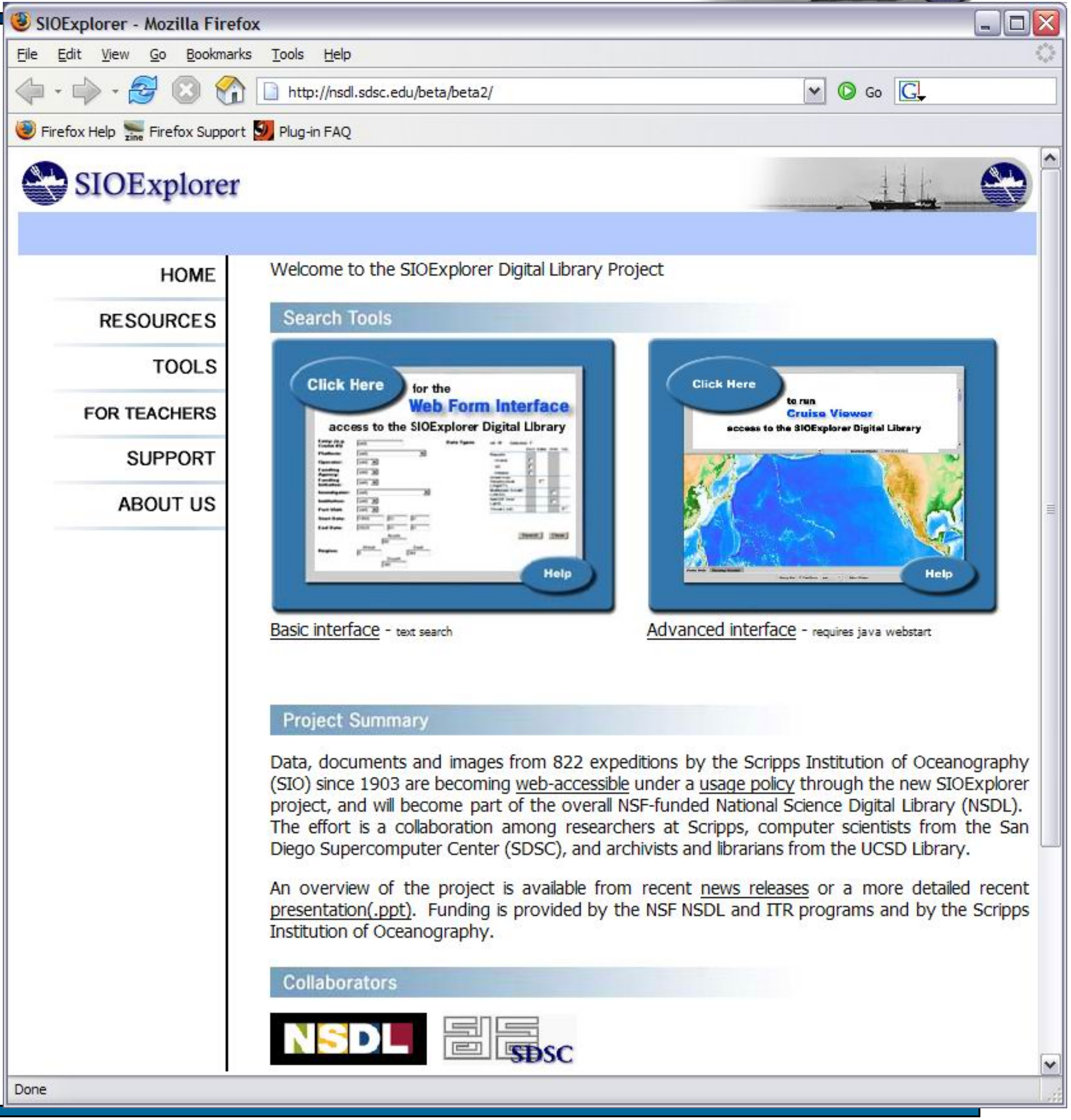
Multi-Institution Testbed
SIO, WHOI, SDSC

SIOExplorer Digital Library

Community access
- Data
- Images
- Documents

647 cruises
150,000 objects
500 GB

Multiple federated collections


SIOExplorer - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://nsdl.sdsc.edu/beta/beta2/
Go
Firefox Help
Firefox Support
Plug-in FAQ
SIOExplorer
HOME
RESOURCES
TOOLS
FOR TEACHERS
SUPPORT
ABOUT US
Click Here
for the
Web Form Interface
access to the SIOExplorer Digital Library
Help
Click Here
to run
Cruise Viewer
access to the SIOExplorer Digital Library
Help
NSDL
SDSC
Done

Welcome to the SIOExplorer Digital Library Project

Search Tools

Basic interface - text search

Advanced interface - requires java webstart

Project Summary

Data, documents and images from 822 expeditions by the Scripps Institution of Oceanography (SIO) since 1903 are becoming web-accessible under a usage policy through the new SIOExplorer project, and will become part of the overall NSF-funded National Science Digital Library (NSDL). The effort is a collaboration among researchers at Scripps, computer scientists from the San Diego Supercomputer Center (SDSC), and archivists and librarians from the UCSD Library.

An overview of the project is available from recent news releases or a more detailed recent presentation(.ppt). Funding is provided by the NSF NSDL and ITR programs and by the Scripps Institution of Oceanography.

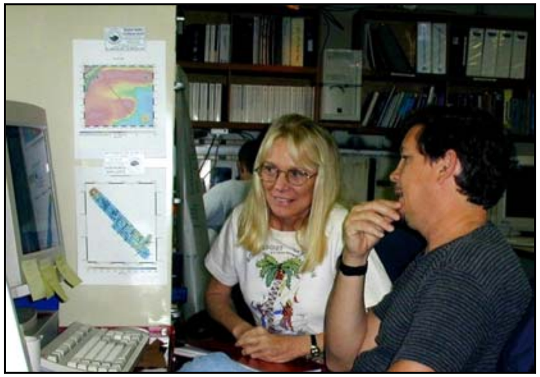
Collaborators

http://SIOExplorer.ucsd.edu

Collection status board

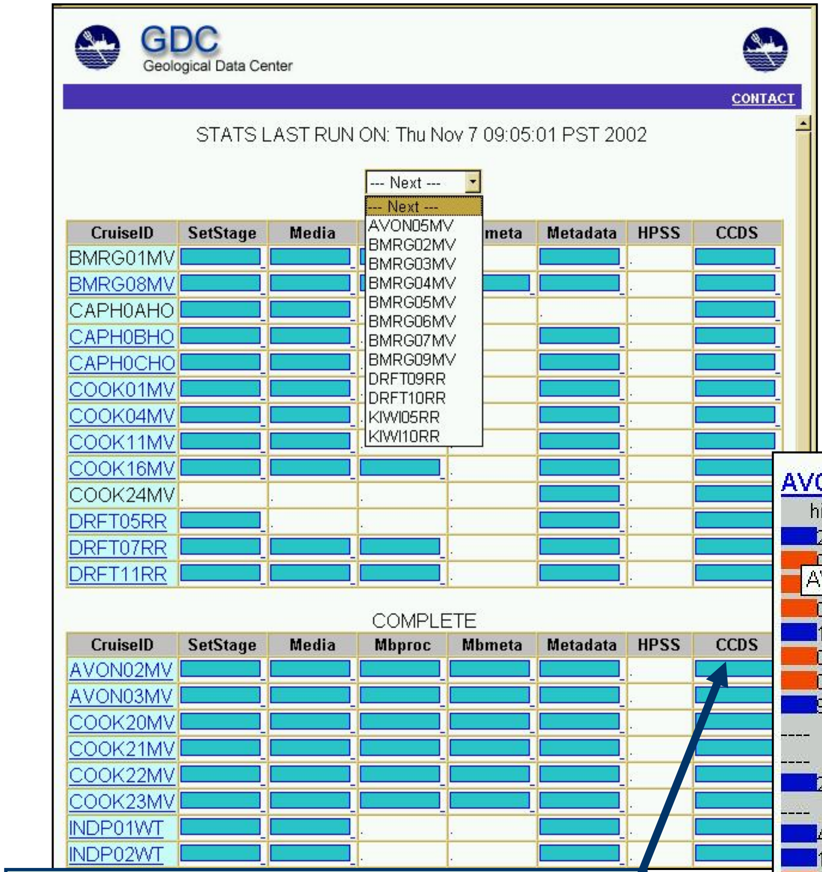
Live on web

Auto-updated



Monitor status of 800 cruises, work in progress

4000 files, 10 GB per cruise



STATS LAST RUN ON: Thu Nov 7 09:05:01 PST 2002

--- Next ---

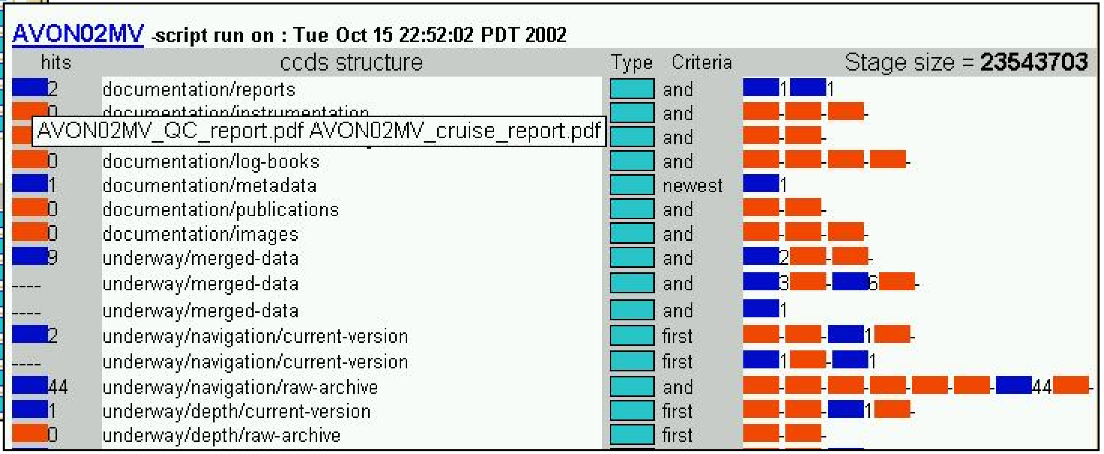
- --- Next ---
- AVON05MV
- BMRG02MV
- BMRG03MV
- BMRG04MV
- BMRG05MV
- BMRG06MV
- BMRG07MV
- BMRG09MV
- DRFT09RR
- DRFT10RR
- KIWI05RR
- KIWI10RR

CruiseID	SetStage	Media		meta	Metadata	HPSS	CCDS
BMRG01MV	▬	▬	▬		▬	.	▬
BMRG08MV	▬	▬	▬	▬	▬	.	▬
CAPH0AHO	▬	▬	.		.	.	▬
CAPH0BHO	▬	▬	.		▬	.	▬
CAPH0CHO	▬	▬	.		▬	.	▬
COOK01MV	▬	▬	.		▬	.	▬
COOK04MV	▬	▬	.		▬	.	▬
COOK11MV	▬	▬	.		▬	.	▬
COOK16MV	▬	▬	▬	.	▬	.	▬
COOK24MV	.	.	.	.	▬	.	▬
DRFT05RR	▬	.	.	.	▬	.	▬
DRFT07RR	▬	▬	▬	.	▬	.	▬
DRFT11RR	▬	▬	▬	.	▬	.	▬

COMPLETE

CruiseID	SetStage	Media	Mbproc	Mbmeta	Metadata	HPSS	CCDS
AVON02MV	▬	▬	▬	▬	▬	.	▬
AVON03MV	▬	▬	▬	▬	▬	.	▬
COOK20MV	▬	▬	▬	▬	▬	.	▬
COOK21MV	▬	▬	▬	▬	▬	.	▬
COOK22MV	▬	▬	▬	▬	▬	.	▬
COOK23MV	▬	▬	▬	▬	▬	.	▬
INDP01WT	▬	▬	.	.	▬	.	▬
INDP02WT	▬	▬	.	.	▬	.	▬

Click for individual cruise status



AVON02MV -script run on : Tue Oct 15 22:52:02 PDT 2002

AVON02MV_QC_report.pdf AVON02MV_cruise_report.pdf

hits	ccds structure	Type	Criteria	Stage size = **23543703**
🟦2	documentation/reports	▬	and	🟦1 🟦1
🟧0	documentation/instrumentation	▬	and	🟧- 🟧- 🟧-
🟧		▬	and	🟧- 🟧-
🟧0	documentation/log-books	▬	and	🟧- 🟧- 🟧- 🟧-
🟦1	documentation/metadata	▬	newest	🟦1
🟧0	documentation/publications	▬	and	🟧- 🟧-
🟧0	documentation/images	▬	and	🟧- 🟧- 🟧-
🟦9	underway/merged-data	▬	and	🟦2 🟧- 🟧-
----	underway/merged-data	▬	and	🟦3 🟧- 🟦6 🟧-
----	underway/merged-data	▬	and	🟦1
🟦2	underway/navigation/current-version	▬	first	🟧- 🟧- 🟦1 🟧-
----	underway/navigation/current-version	▬	first	🟦1 🟧- 🟦1
🟦44	underway/navigation/raw-archive	▬	and	🟧- 🟧- 🟧- 🟧- 🟧- 🟧- 🟦44 🟧-
🟦1	underway/depth/current-version	▬	first	🟧- 🟧- 🟦1 🟧-
🟧0	underway/depth/raw-archive	▬	first	🟧- 🟧-

Issue for future use:
Access to complete cruise collections

- Current practice hit-or-miss
  - Only selected data streams archived
- Cyberinfrastructure allows comprehensive solution
  - Auto-harvesting and archiving
    - Data and metadata

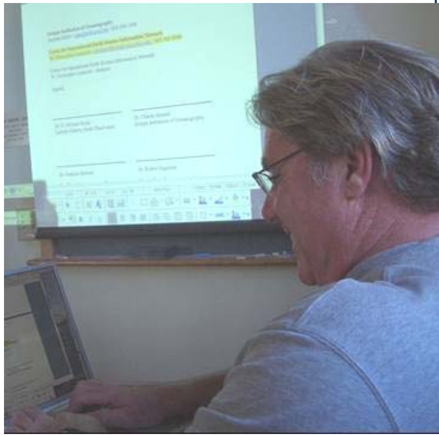
Claim:
Very little additional cost to archive everything

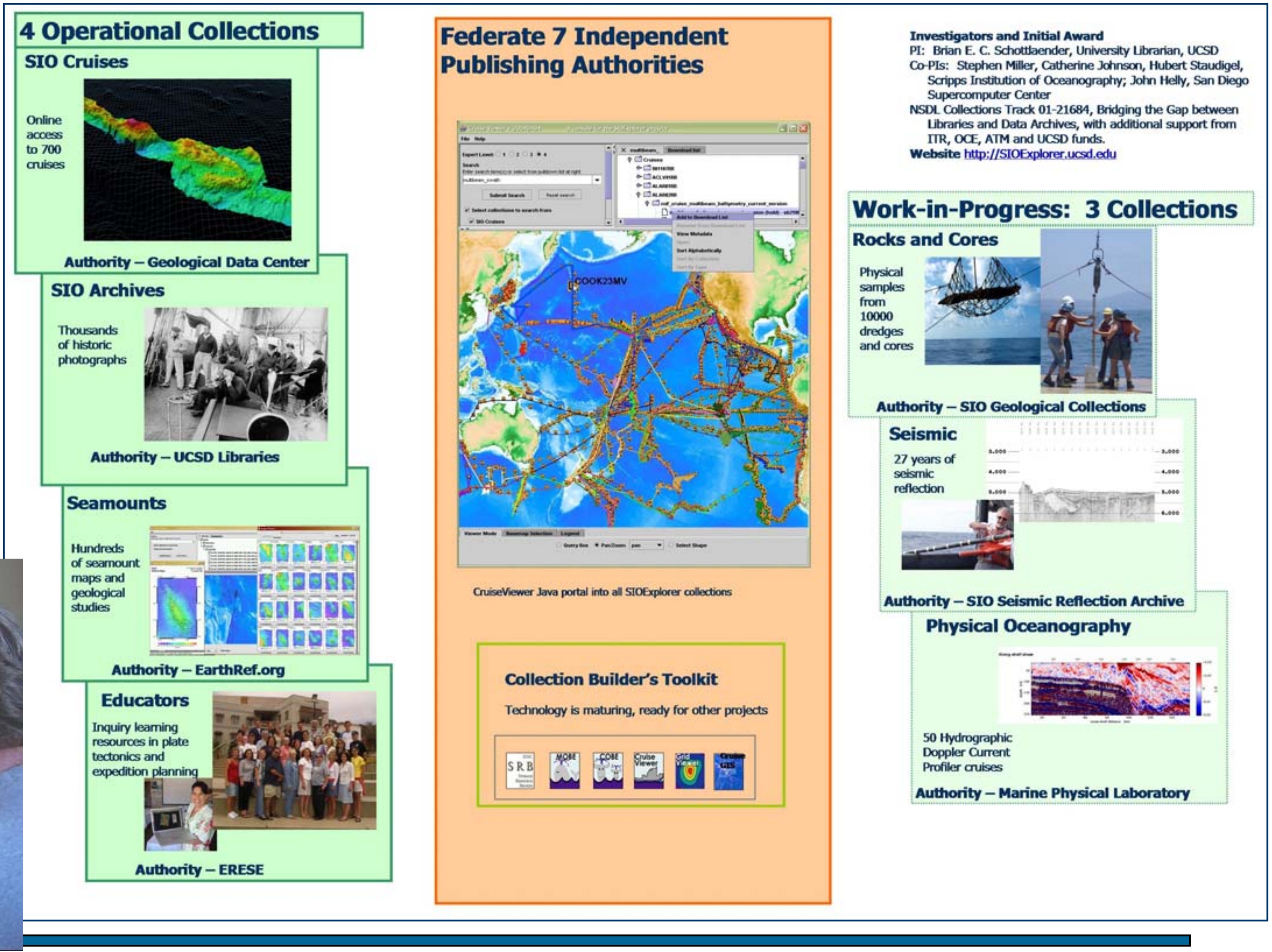
Design to Overcome Project Barriers

Build scalable digital library

Federate independent authorities

4 Operational collections
3 Work-in-progress





4 Operational Collections

SIO Cruises

Online access to 700 cruises

Authority – Geological Data Center

SIO Archives

Thousands of historic photographs

Authority – UCSD Libraries

Seamounts

Hundreds of seamount maps and geological studies

Authority – EarthRef.org

Educators

Inquiry learning resources in plate tectonics and expedition planning

Authority – ERESE

Federate 7 Independent Publishing Authorities

CruiseViewer Java portal into all SIOExplorer collections

Collection Builder's Toolkit

Technology is maturing, ready for other projects

Investigators and Initial Award
PI: Brian E. C. Schottlaender, University Librarian, UCSD
Co-PIs: Stephen Miller, Catherine Johnson, Hubert Staudigel, Scripps Institution of Oceanography; John Helly, San Diego Supercomputer Center
NSDL Collections Track 01-21684, Bridging the Gap between Libraries and Data Archives, with additional support from ITR, OCE, ATM and UCSD funds.
Website http://SIOExplorer.ucsd.edu

Work-in-Progress: 3 Collections

Rocks and Cores

Physical samples from 10000 dredges and cores

Authority – SIO Geological Collections

Seismic

27 years of seismic reflection

Authority – SIO Seismic Reflection Archive

Physical Oceanography

50 Hydrographic Doppler Current Profiler cruises

Authority – Marine Physical Laboratory

Multiple access methods

Google

No interface

Just type name of cruise

Basic web form

Text-based search for experts

Java CruiseViewer

Full graphical search

Web services

Computer-to-computer

Enable next generation interoperability

Java CruiseViewer

Full graphical search

- All capabilities
- Any combination of collections

Metadata

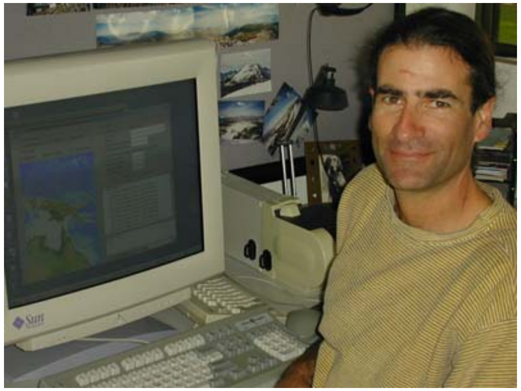
- Oracle or PostgreSQL

Data

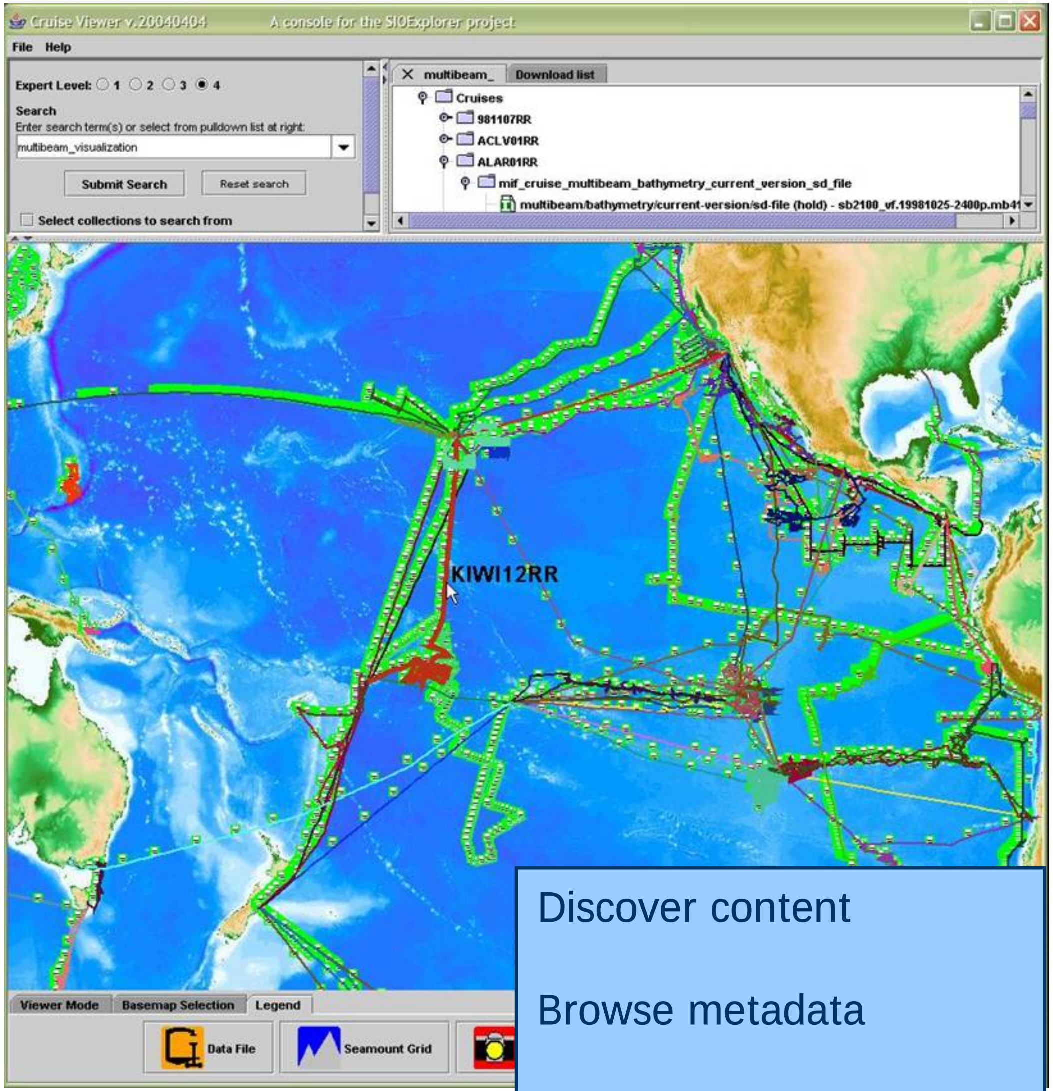
- Storage Resource Broker

User

- Graphical search
- Keyword search



Don Sutton, SDSC



Cruise Viewer v.20040404 A console for the SIOExplorer project

File Help

Expert Level: ○ 1 ○ 2 ○ 3 ● 4

Search

Enter search term(s) or select from pulldown list at right:

multibeam_visualization

Submit Search Reset search

□ Select collections to search from

multibeam_ | Download list

- Cruises
  - 981107RR
  - ACLV01RR
  - ALAR01RR
    - mif_cruise_multibeam_bathymetry_current_version_sd_file
      - multibeam/bathymetry/current-version/sd-file (hold) - sb2100_vf.19981025-2400p.mb41

Search results for visualization objects

Discover content

Browse metadata

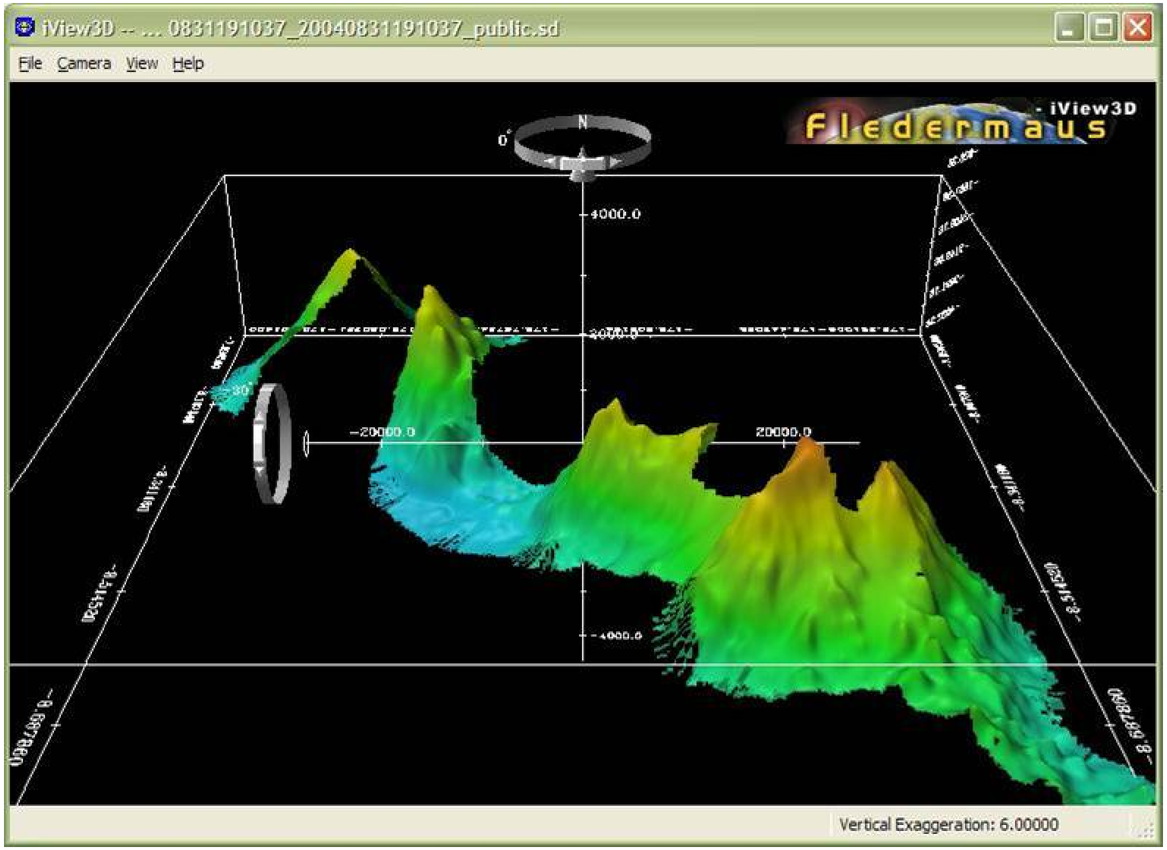
View or download objects

Launch visualization experiences

Visualization of multibeam seafloor mapping swath sonar data

300 cruises since 1982
20-km wide swaths

Sonar quality control
Geological research
Education



Vertical Exaggeration: 6.00000

Download free viewer

http://www.ivs.unb.ca/products/iview3d/

Broader Impact with ERESE National Teachers Workshops

Enduring Resources for Earth Science Education

Two-week summer workshops
2004 and 2005

Build inquiry-driven learning experiences

THE NATIONAL SCIENCE DIGITAL LIBRARY

Other organizations using mtf technology

CUAHSI Consortium of Universities for Advanced Hydrologic Science, Inc.

Major technology co-development

95 institutional members

WHOI – DIGARCH Multi-Institution Testbed project

Bob Detrick

CCOM/UNH cruise and multibeam archives

Jim Case, Larry Mayer

MBARI – Monterey Bay Aquarium Research Institute collection building in progress

Dave Caress, Andrew Chase

SOEST/HAWAII – April 4-26, 2005 realtime digital library testing R/V Kilo Moana

NIWA – Digital-Library-in-a-Box tested on R/V Tangaroa in New Zealand

John Helly, Don Robertson

Arctic DMS - Data Management System under development

Margo Edwards (Hawaii), Dawn Wright (Oregon State)

Multi-Institution Testbed for Scalable Digital Archiving

Extend SIOExplorer approach to WHOI

Integrate SIO, SDSC and WHOI tools and data

30 years of WHOI cruise data

4098 Alvin submersible dives

Jason ROV surveys (200 DVD per cruise)

Results from 1600 NSF awards online

Project Challenges

Auto-harvest data, metadata

“Shoe-box archives” only

prior to 2002

Build distributed digital library

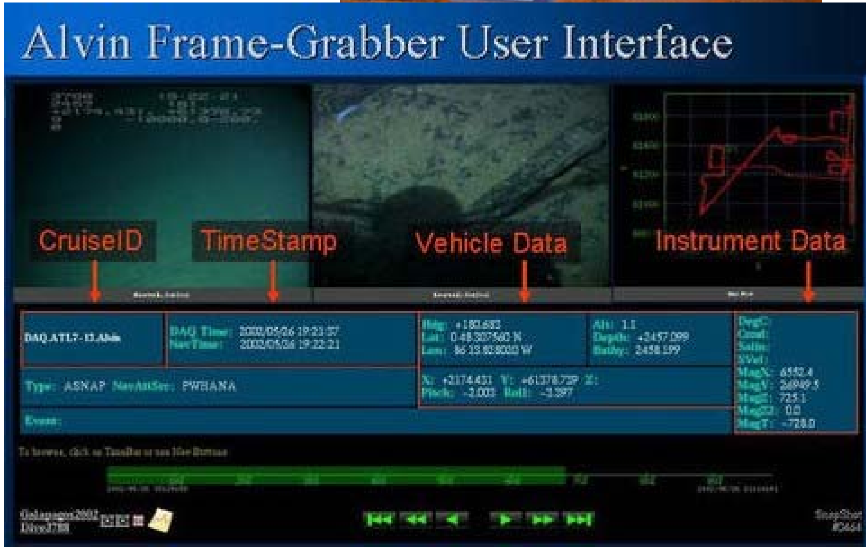
Both institutions

Ships and submersibles

Extend WHOI data exploration tools

Persistent digital library objects

Interoperability across institutions



Alvin Frame-Grabber User Interface

DAQ.ATL7-13.Alvin

DAQ Time: 2002/05/26 19:21:37
NavTime: 2002/05/26 19:22:21

Hdg: +180.682
Lat: 0 48.307560 N
Lon: 86 13.828020 W

Alt: 1.1
Depth: +2457.099
Bathy: 2458.199

Type: ASNAP NavAttSrc: PWHANA

X: +2174.431 Y: +61378.729 Z:
Pitch: -2.003 Roll: -3.297

Event:

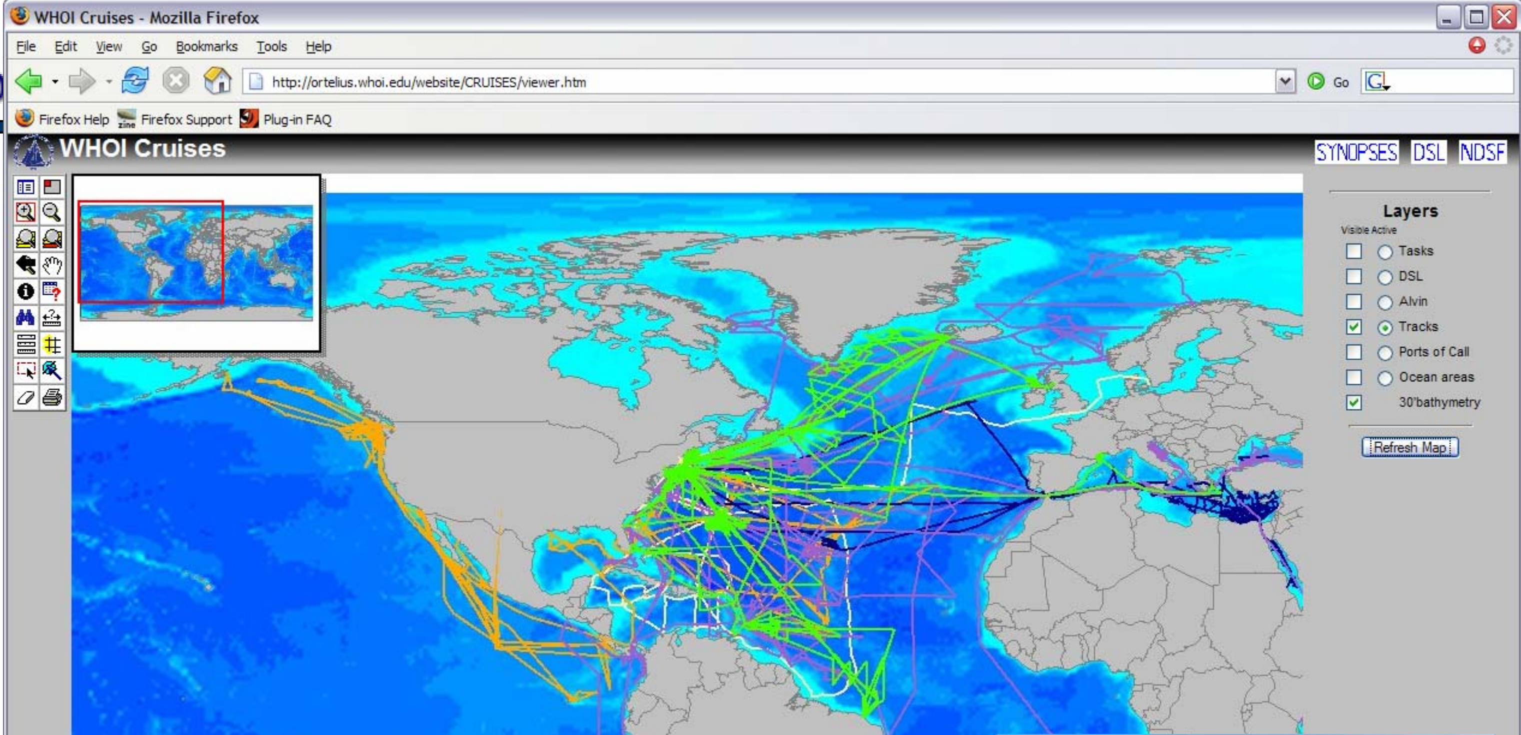
DegC:
Cond:
Salin:
SVel:
MagX: 6552.4
MagY: 26949.5
MagZ: 725.1
MagZ2: 0.0
MagT: -728.0

To browse, click on TimeBar or use Nav Buttons

Galapagos2002
Dive3788

SnapShot
#0464

WHOI cruises


WHOI Cruises - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://ortelius.whoi.edu/website/CRUISES/viewer.htm
Go
Firefox Help
Firefox Support
Plug-in FAQ
WHOI Cruises
SYNOPSES
DSL
NDSF

Layers

Visible Active

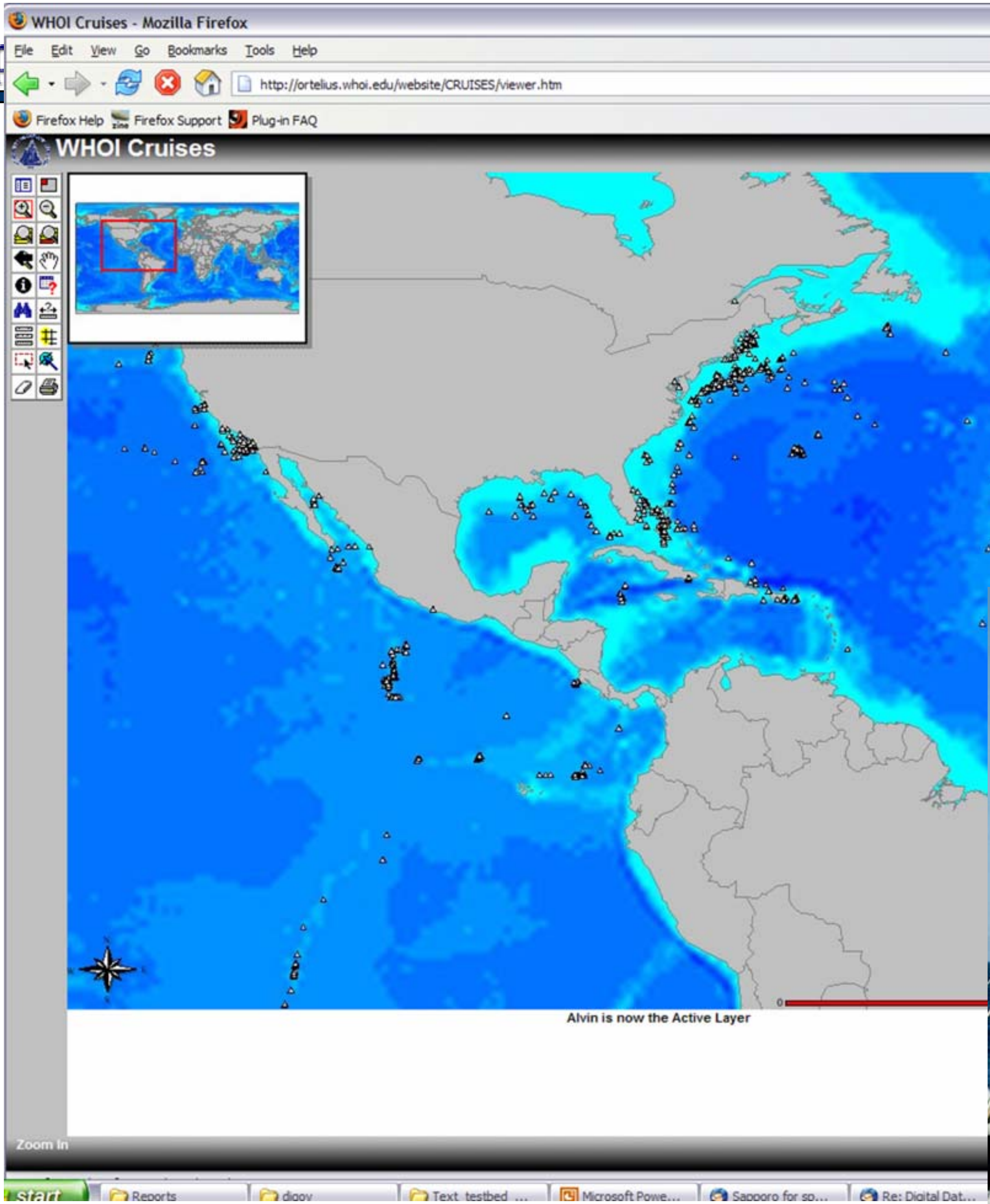
- □ ○ Tasks
- □ ○ DSL
- □ ○ Alvin
- ☑ ● Tracks
- □ ○ Ports of Call
- □ ○ Ocean areas
- ☑ 30'bathymetry

Refresh Map

800 cruises since 1930

4098 Alvin dives

Since June 26, 1964


WHOI Cruises - Mozilla Firefox

http://ortelius.whoi.edu/website/CRUISES/viewer.htm

WHOI Cruises

Layers

Visible Active

- □ ○ Tasks
- □ ○ DSL
- ☑ ● Alvin
- □ ○ Tracks
- □ ○ Ports of Call
- □ ○ Ocean areas
- ☑ 5'bathymetry

Alvin is now the Active Layer