U.S. JGOFS Data Management
a retrospective

Cyndy Chandler
U.S. JGOFS Data Management Office
26 May 2005

Data Management / Data Policy Discussion
Washington, DC

Topics in today's presentation:

- Introduction
  - What is U.S. JGOFS?
  - What is the DMO?
- Lessons Learned
  - U.S. JGOFS data server
    - JGOFS d-DBMS and user interface
    - Live Access Server

What is U.S. JGOFS?
U.S. Joint Global Ocean Flux Study

- part of multinational JGOFS
- U.S. Global Change Research Program (US GCRP), Scientific Committee on Oceanic Research (SCOR), and International Geosphere-Biosphere Programme (IGBP)
- long term (U.S. 1989-2005)
- multidisciplinary (bio, chem, PO, geology)
- process studies (U.S. 1989-1998), time-series, global surveys, synthesis and modeling, data management
- investigate ocean carbon flux

U.S. JGOFS
Data Management Office (DMO)

- formed in 1988 specifically to meet needs of U.S. JGOFS
- assist PIs to submit their data to DMO
- ongoing quality control of data
- develop and maintain simple, reliable interface to program data
- provide timely, easy access to project results
- collaborate with other program DMO
- publish U.S. JGOFS data reports
- plan final archive of U.S. JGOFS information

Basic Principles of Data Management

from 1988 JGOFS Working Group on Data Management

- scientists will generate data in a format useful for their needs
- oceanographic data sets are best organized in terms of metadata (temporal and geographical)
- data managers should avoid use of coded data values
- users should be able to obtain all the data they require from one source and in a consistent format
- data interchange formats should be designed for the convenience of scientific users

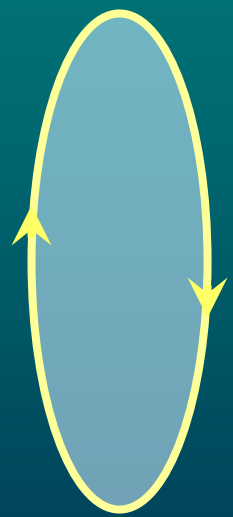
Additional Guidelines

- metadata is critical and therefore mandatory
- data managers should maintain awareness of emerging standards and strive for compliance
- whatever interface is used to provide access to the data, it's still important to provide subset and download capability
- data management systems must be dynamic - balancing tension between existing and new technologies

Basic Components of a Data Management System

first four components are active throughout the program

- data acquisition
  - from variety of sources
- quality assurance
- data publication
- synthesis and modeling



- archive

from the American Heritage Dictionary of the English Language

lesson defined . . .

1 : a passage from sacred writings read in a service of worship

2 a : a piece of instruction **b :** a reading or exercise to be studied by a pupil **c :** a division of a course of instruction

3 a : something learned by study or experience
b : an instructive example

4 : an edifying example or experience

5 : a reprimand

Lessons Learned . . .

The first lesson learned . . .

is a meta lesson . . .

All the lessons learned take on enhanced meaning when applied to science programs of increasing size and complexity.

Lessons Learned . . .

Lessons Learned . . .

- technology is good; people are more important
  - ✓ diverse range of expertise and personalities
    - designers, programmers, data managers
  - ✓ someone with authority to set overall vision
- ✓ qualified staff to make effective use of technology
- ✓ guidance from advisory committee which includes active investigators

U.S. JGOFS DMO personnel

- David Glover (director)
- Cyndy Chandler (manager)

- Previous staff members
  - Christine Hammond (manager)
  - George Heimerdinger (data specialist)
  - David Schneider (data specialist)
  - Jeff Dusenberry (data specialist)

Lessons Learned . . .

- develop a data policy, publicize and follow it
  - ✓ guidance from steering committee
  - ✓ consent from participating investigators
  - ✓ reiterate at conferences and workshops
  - ✓ encourage compliance
  - ✓ agree on method of enforcement (used as last resort)

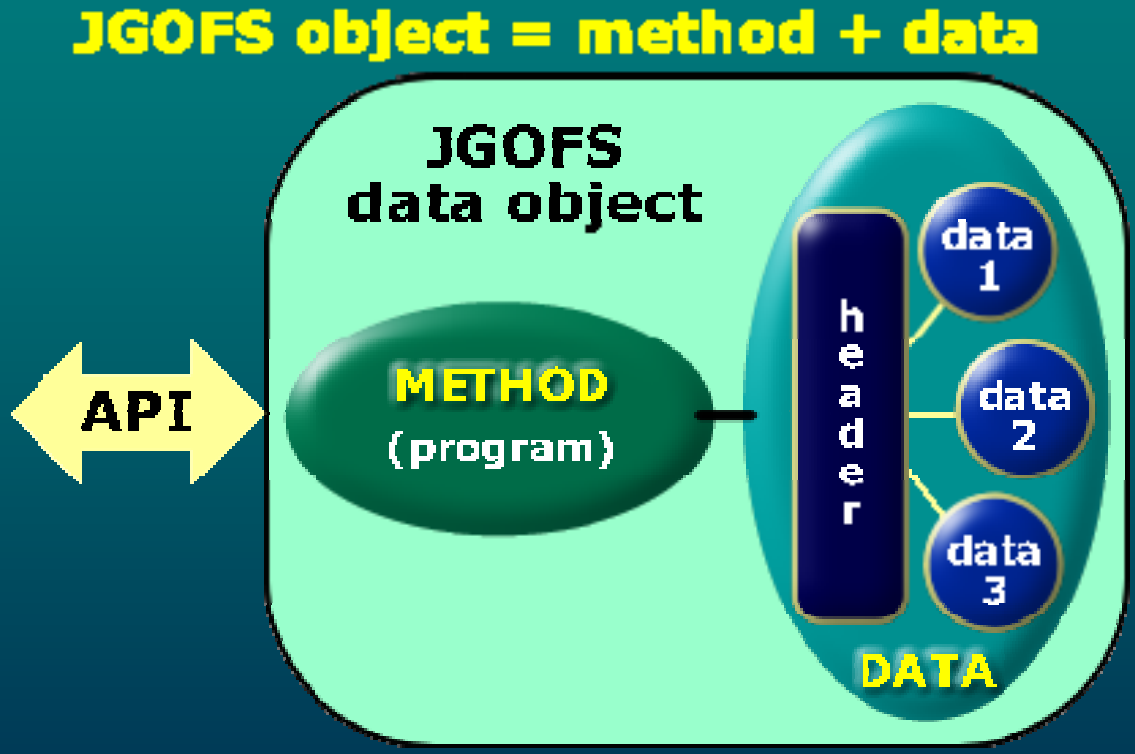
Lessons Learned . . .

- establish protocols at program start with mechanism for adaptation when necessary
  - ✓ sampling methodologies
  - ✓ naming conventions
    - ❖ parameter dictionary, controlled vocabulary
    - ❖ XML schema, thesauri, ontologies
  - ✓ units of measurement

Lessons Learned . . .

- facilitate contribution of data to collection
  - ✓ compile an inventory of expected results
  - ✓ publish and maintain the inventory
  - ✓ remind investigators of opportunities to contribute data and results to the growing inventory
  - ✓ review procedure at conferences and workshops
  - ✓ accept all formats of data
  - ✓ work with investigators to complete metadata records

JGOFS distributed database management system

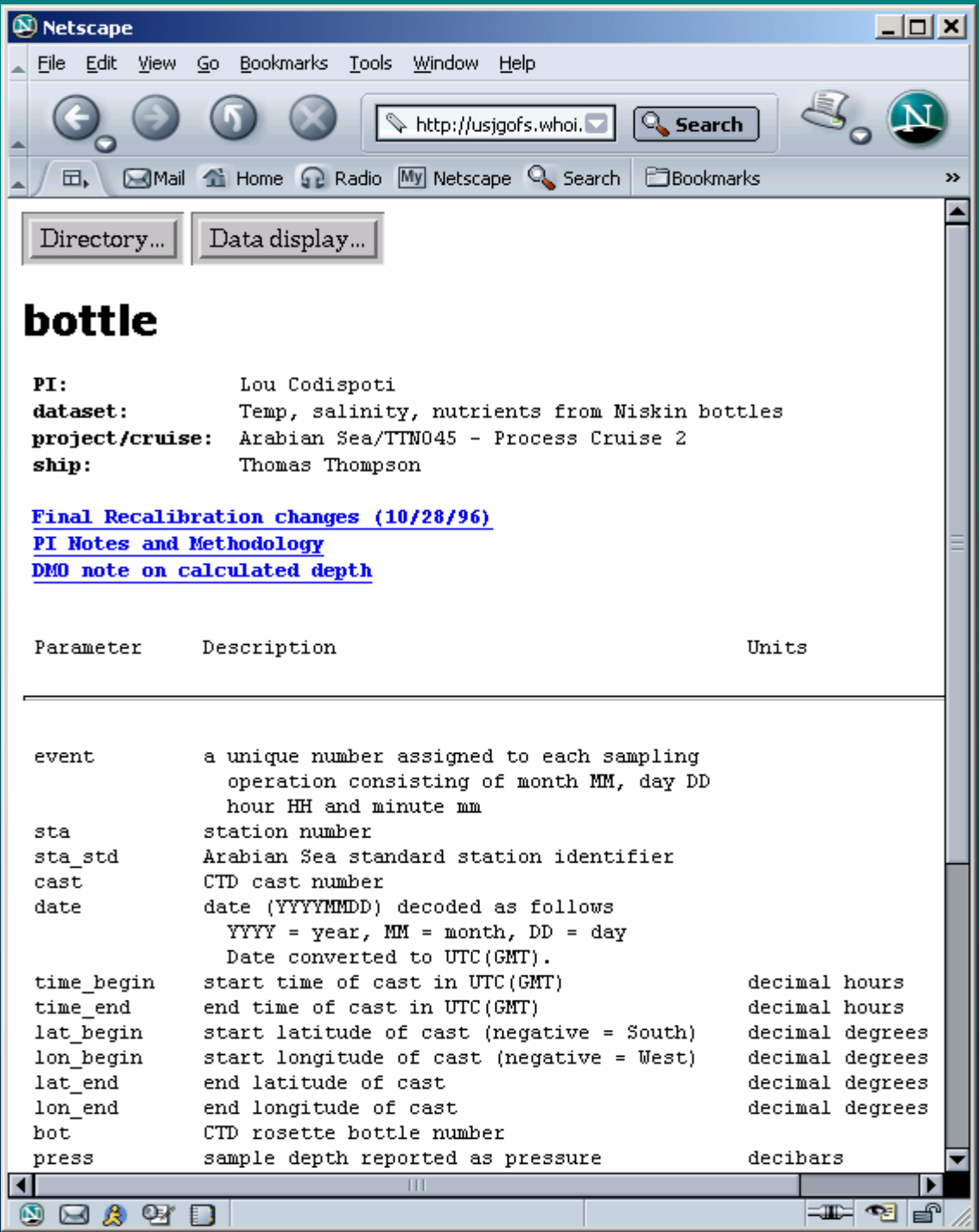


U.S. JGOFS DMO accepted any format data from the field study investigators and used methods to locate and translate the data objects

Lessons Learned . . .

- metadata is of critical importance
  - accurate, complete, available with data
  - monitor emerging standards
  - define minimal metadata requirements
  - standards-compliant solutions where possible
  - complete metadata record enables reuse of data

- metadata assembly is time consuming, but is the key to enabling secondary (reuse) of data


Netscape
File Edit View Go Bookmarks Tools Window Help
http://usjgofs.whoi.
Search
Mail Home Radio Netscape Search Bookmarks
Directory... Data display...
bottle
PI: Lou Codispoti
dataset: Temp, salinity, nutrients from Niskin bottles
project/cruise: Arabian Sea/TTN045 - Process Cruise 2
ship: Thomas Thompson
Final Recalibration changes (10/28/96)
PI Notes and Methodology
DMO note on calculated depth
Parameter Description Units
event a unique number assigned to each sampling operation consisting of month MM, day DD hour HH and minute mm
sta station number
sta_std Arabian Sea standard station identifier
cast CTD cast number
date date (YYYYMMDD) decoded as follows YYYY = year, MM = month, DD = day Date converted to UTC(GMT).
time_begin start time of cast in UTC(GMT) decimal hours
time_end end time of cast in UTC(GMT) decimal hours
lat_begin start latitude of cast (negative = South) decimal degrees
lon_begin start longitude of cast (negative = West) decimal degrees
lat_end end latitude of cast decimal degrees
lon_end end longitude of cast decimal degrees
bot CTD rosette bottle number
press sample depth reported as pressure decibars

metadata

Lessons Learned . . .

- quality assurance is an ongoing process
  - intense QA process during initial acquisition and ingestion into data system
  - problems discovered as data are utilized by others
    - insufficient or inaccurate metadata
  - process of data product synthesis becomes a valuable diagnostic tool for improving data quality

Lessons Learned . . .

- begin synthesis early
  - ✓ do not wait until all the data has been collected
  - ✓ synthesized products greatly enhance the growing collection of data

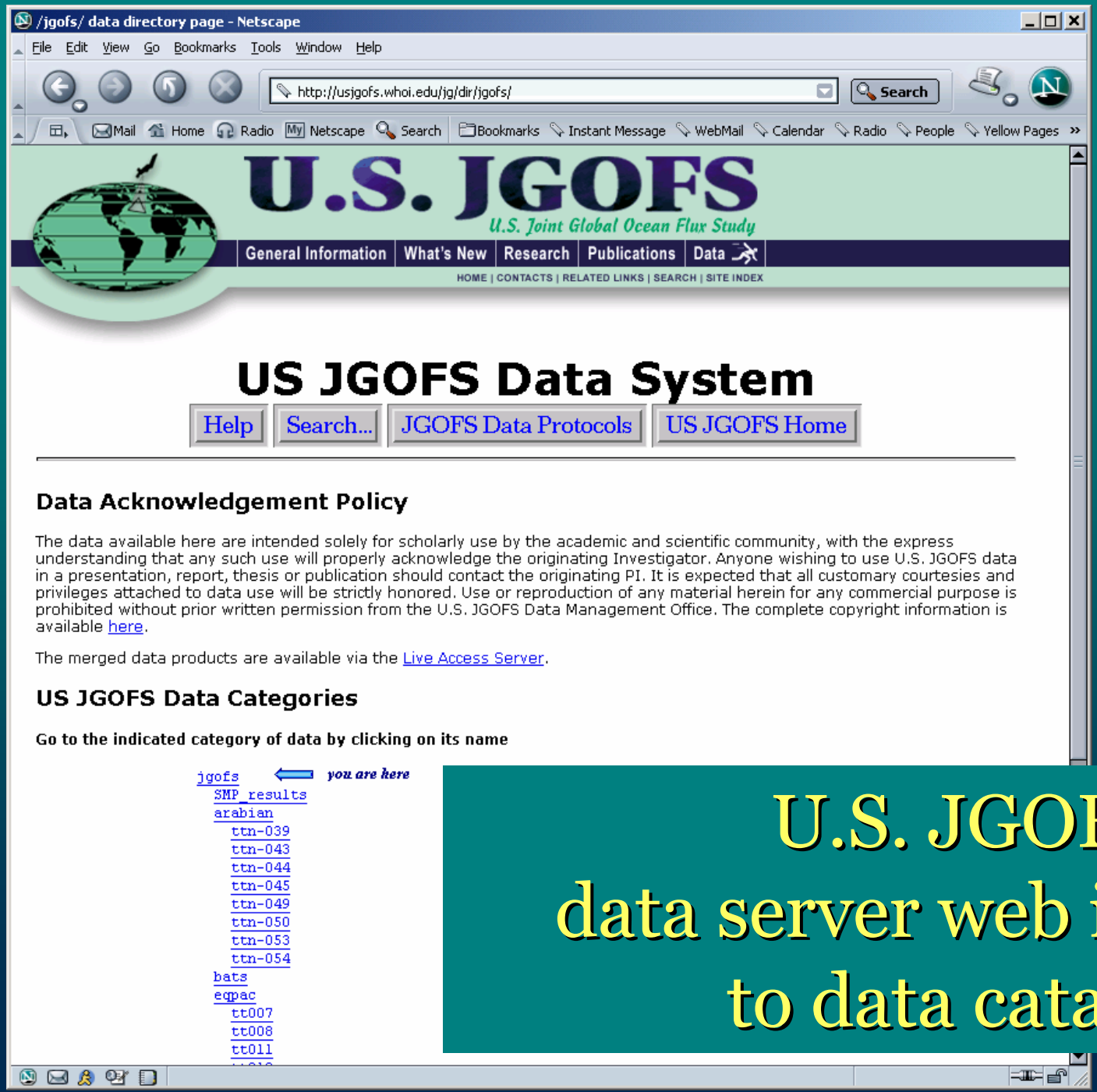
Lessons Learned . . .

- provide timely, easy access to project results
- develop and maintain simple, reliable interface to data collection
  - single interface to entire data collection
  - balance tension between new innovative technologies and existing stable implementations
  - if an interface is broken, it doesn't matter how great the original concept was

skip data system tech details

U.S. JGOFS Data Server

- JGOFS distributed database management system (d-DBMS) used for field data
- Live Access Server (LAS) used for gridded, synthesis and model results

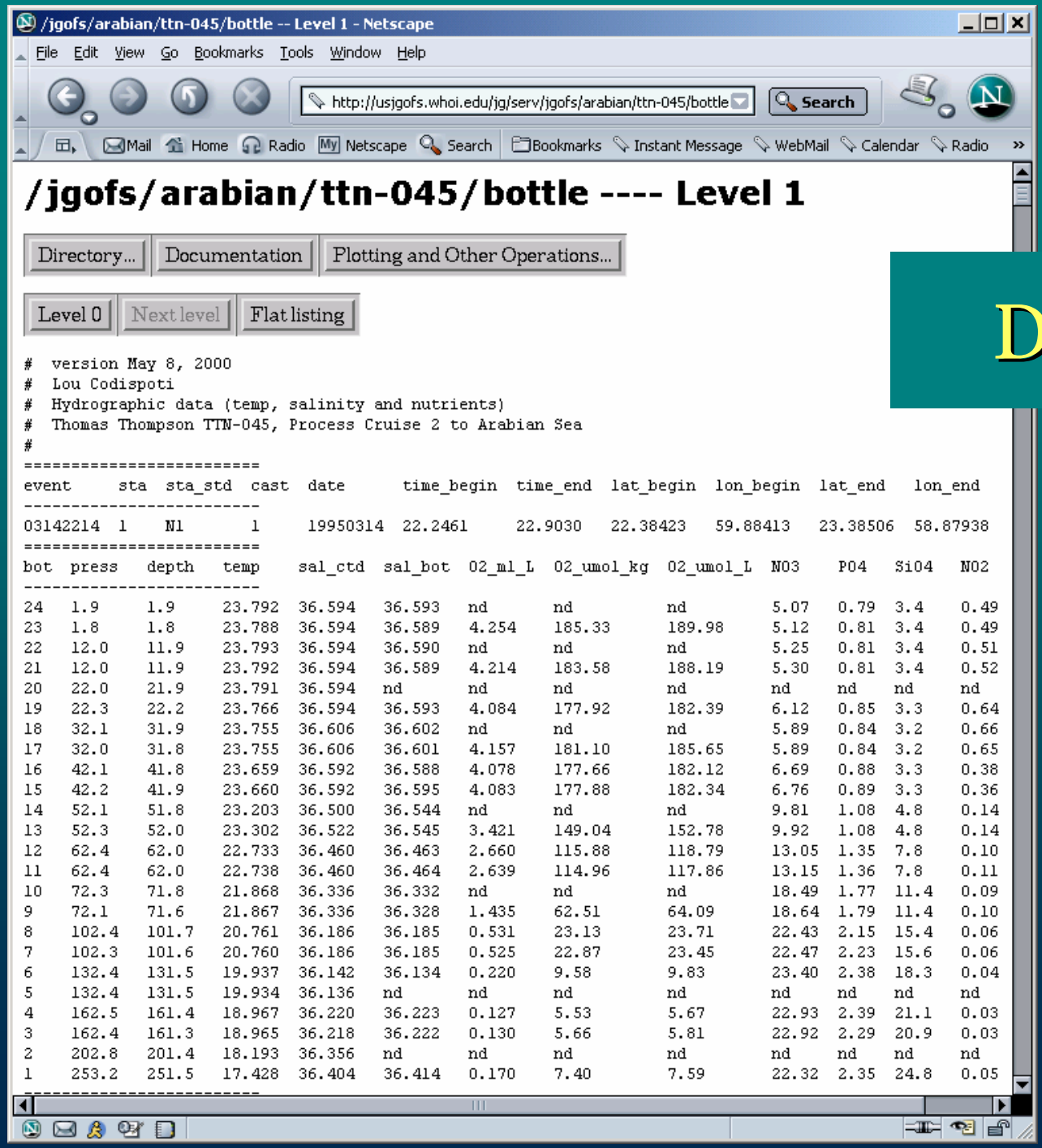


U.S. JGOFS data server web interface to data catalog

Data

JGOFS Distributed Database Management System (d-DBMS)

- distributed, object-oriented system
- originally developed by Glenn Flierl, James Bishop, Satish Paranjpe, David Glover
- supports multidisciplinary, multi-institutional data acquisition project
- multiple data storage formats and locations
- data interpreted by ‘methods’

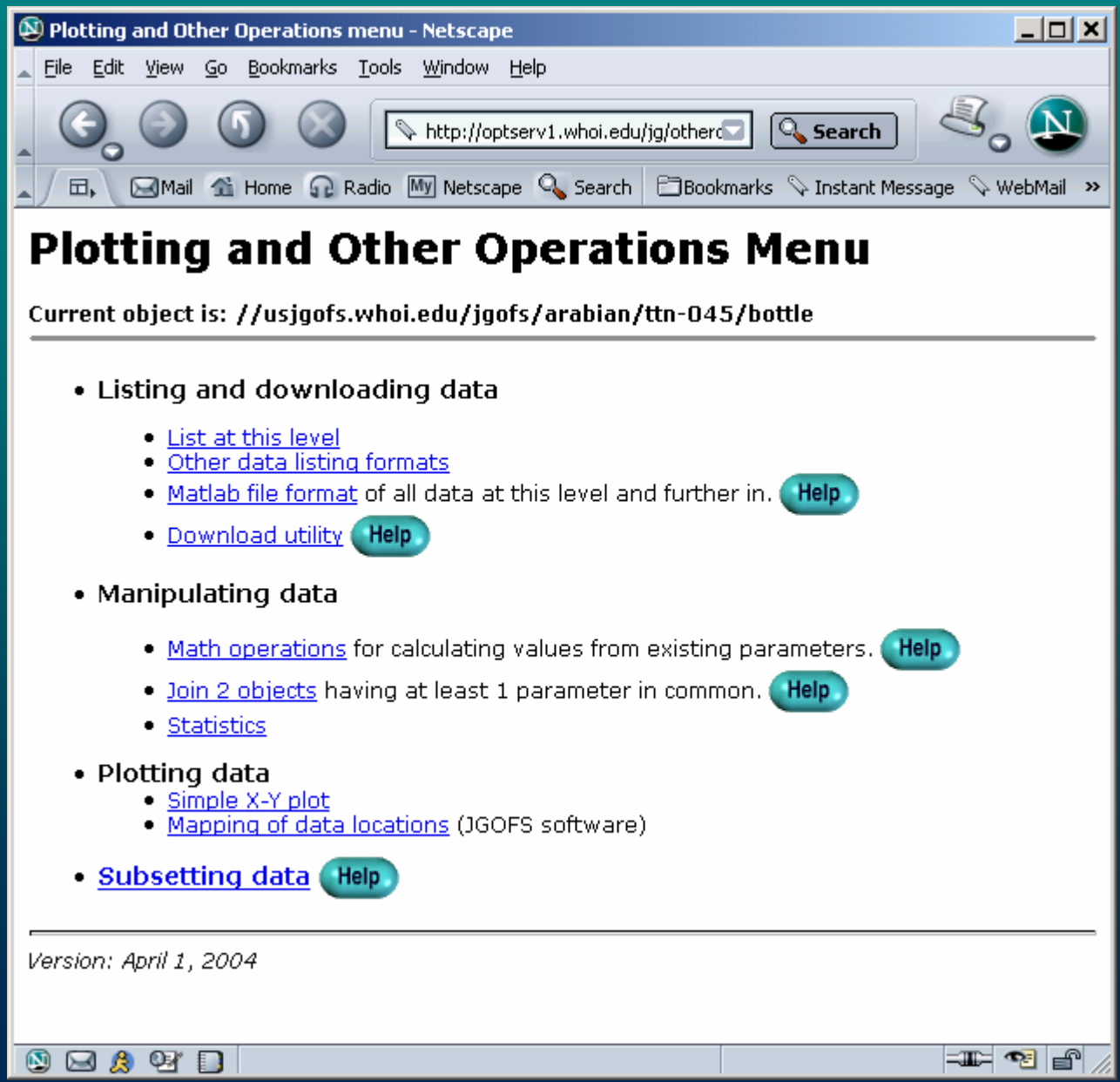


Data Listing

- ✓ select dataset
- ✓ select variable (columns)
- ✓ select range (rows)
- ✓ view metadata

Data

subset, plot, download data



Live Access Server Interface

provides access to synthesis and model results

Live Access Server ~ LAS

- LAS Development Team (original)
- Steve Hankin
- Jon Callahan
- Joe Sirott

- located at UW JISAO/NOAA-PMEL

  University of Washington's Joint Institute for the Study of the Atmosphere and Ocean and NOAA Pacific Marine Environmental Laboratory

Live Access Server Interface

- configurable Web server
- data and metadata interface
- provides access to geo-referenced scientific data
- presents distributed data sets as a unified virtual data base (DODS/OPeNDAP)
- uses Ferret as the default visualization application
- visualize data with on-the-fly graphics
- request custom subsets of variables in a variety of file formats
- access background reference material (metadata)
- compare variables from distributed sources

LAS enables a data server to ...

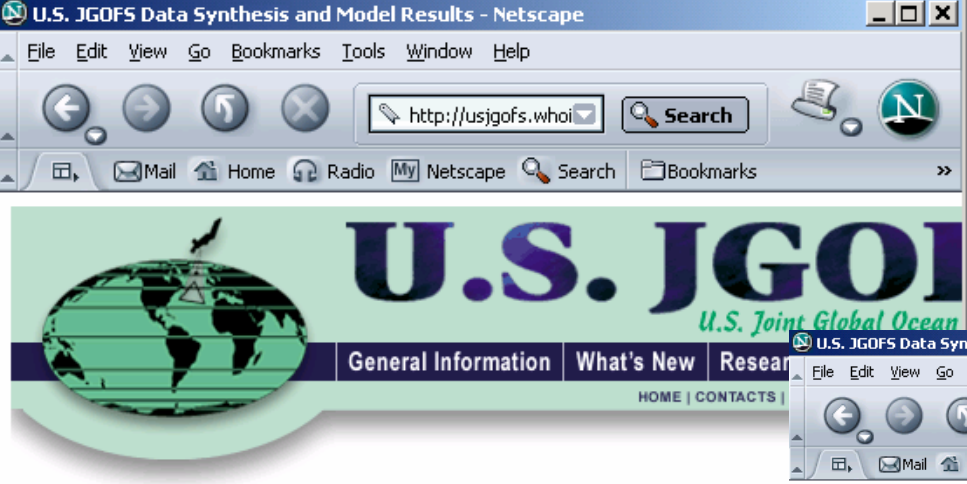
- unify access to multiple types of data in a single interface
- create thematic data collections from distributed data sources
- offer derived products on the fly
- offer variety of visualization styles
  - customized for the data

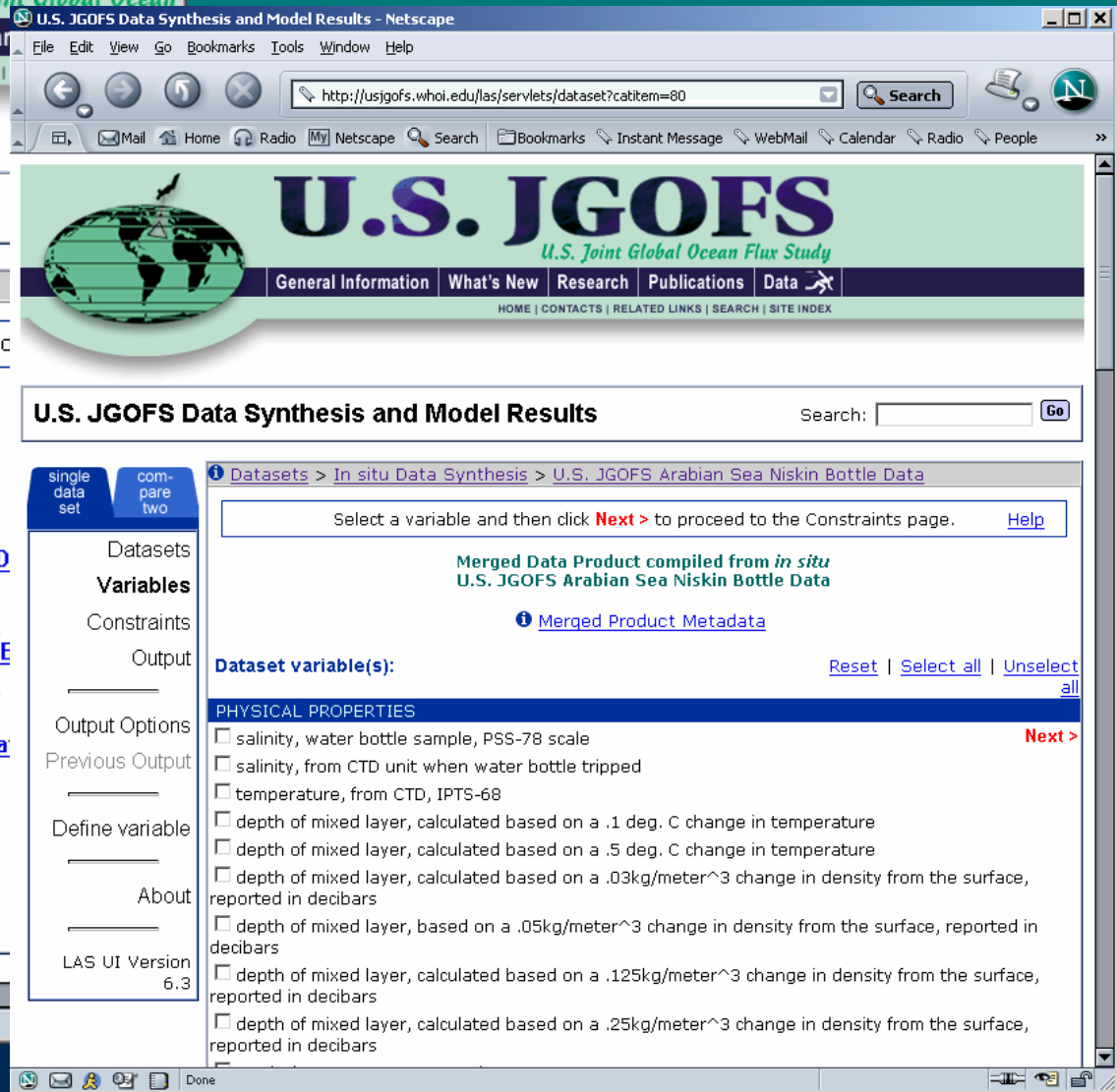
U.S. JGOFS LAS

- MySQL database of netCDF and JGOFS format data objects
- interface to project data and metadata
- data sub-selection (selections, projections)
- multi-variable support
- gridded vs. in-situ data differencing
- multiple views
  - (property-property, depth horizon, cruise tracks, overplots)
- multiple products (ps, gif, text, NetCDF)

LAS v6

select dataset and variables





LAS v6 constraints

- ✓ select dataset
- ✓ select variable
- ➢ set constraints

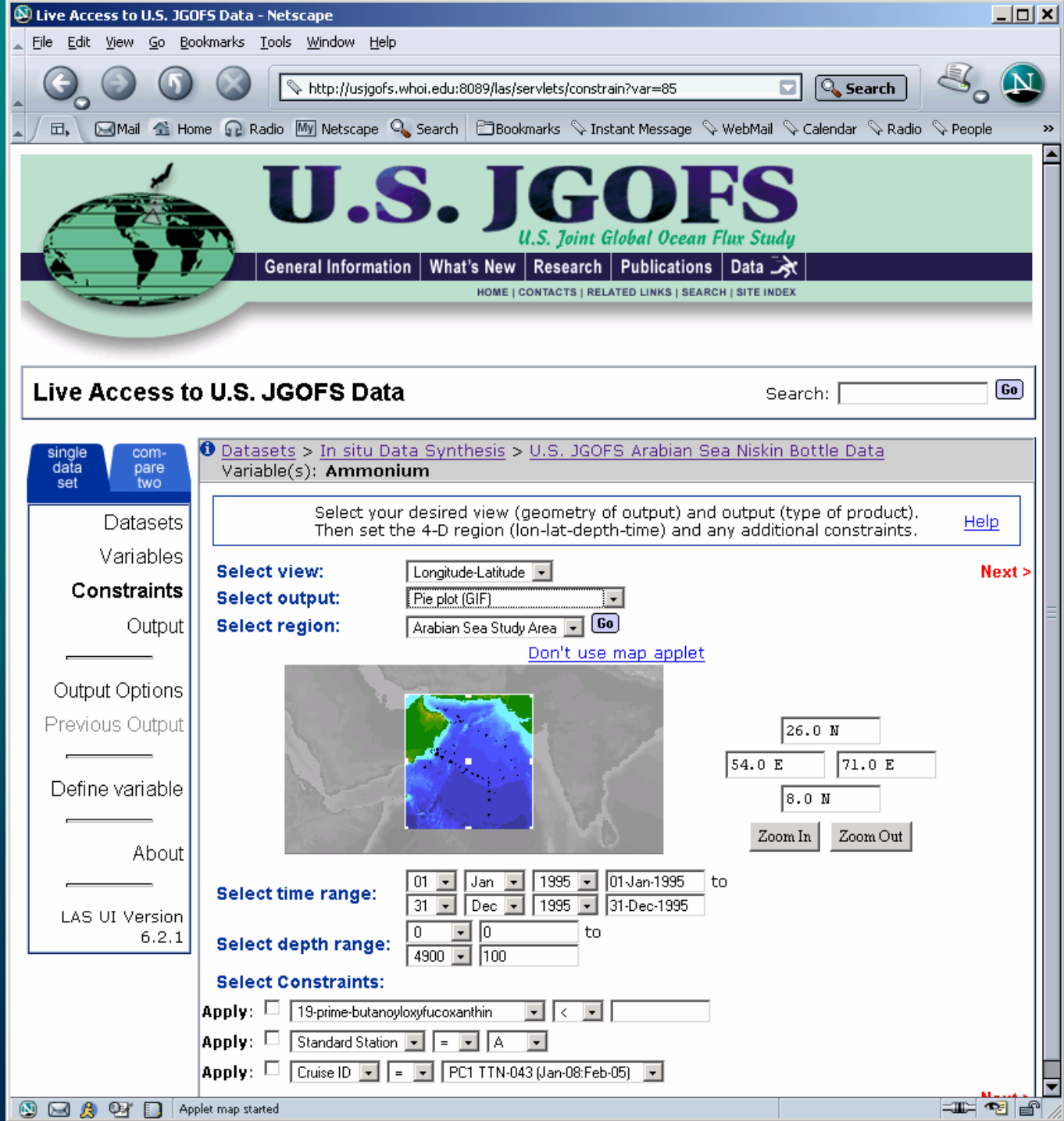
select view (XYZT)

select output type

selections:

- lat/lon
- time
- depth range

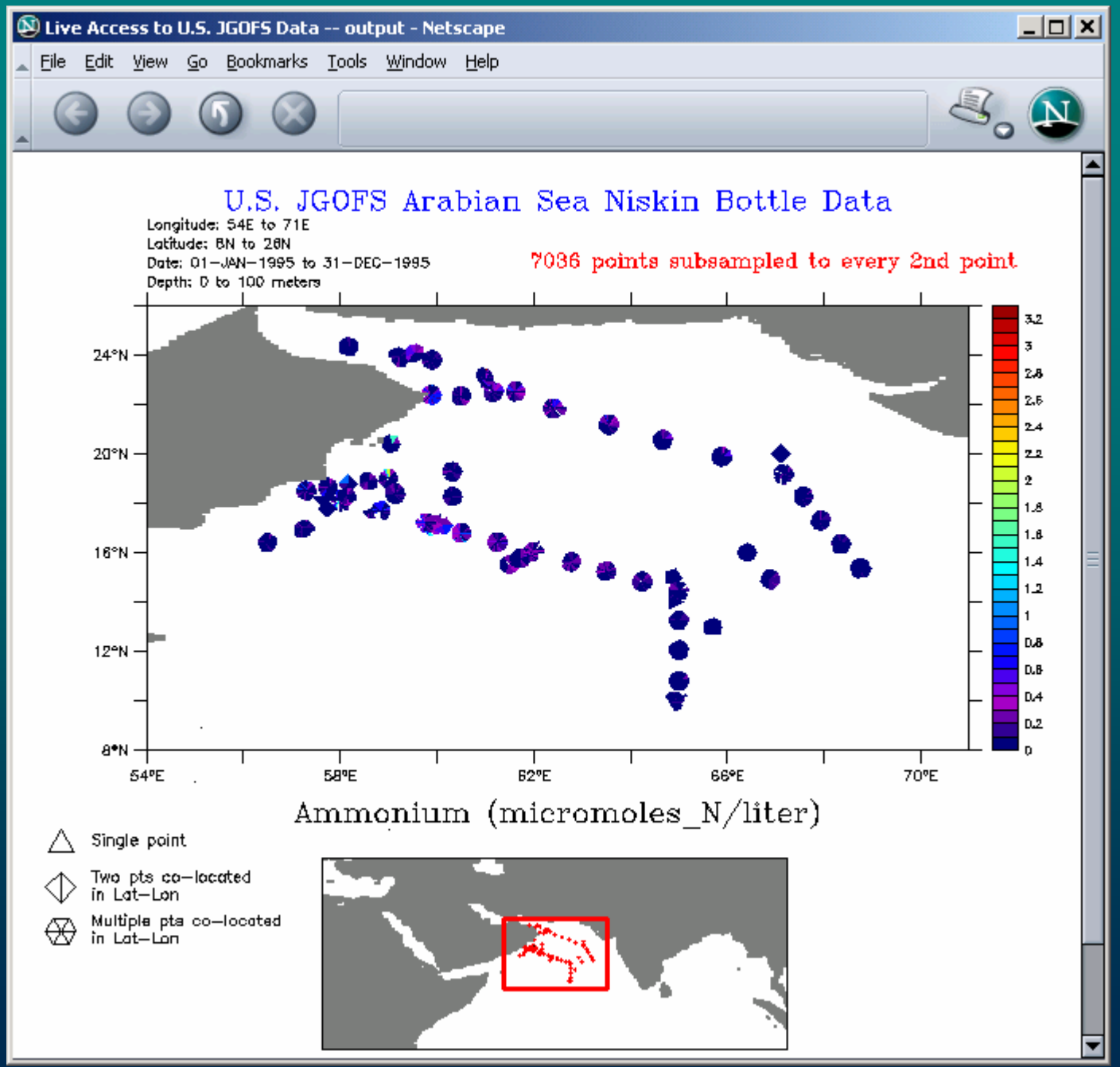
26 May 2005



- ✓ select dataset
- ✓ select variable
- ✓ set constraints
- ➢ output

LAS

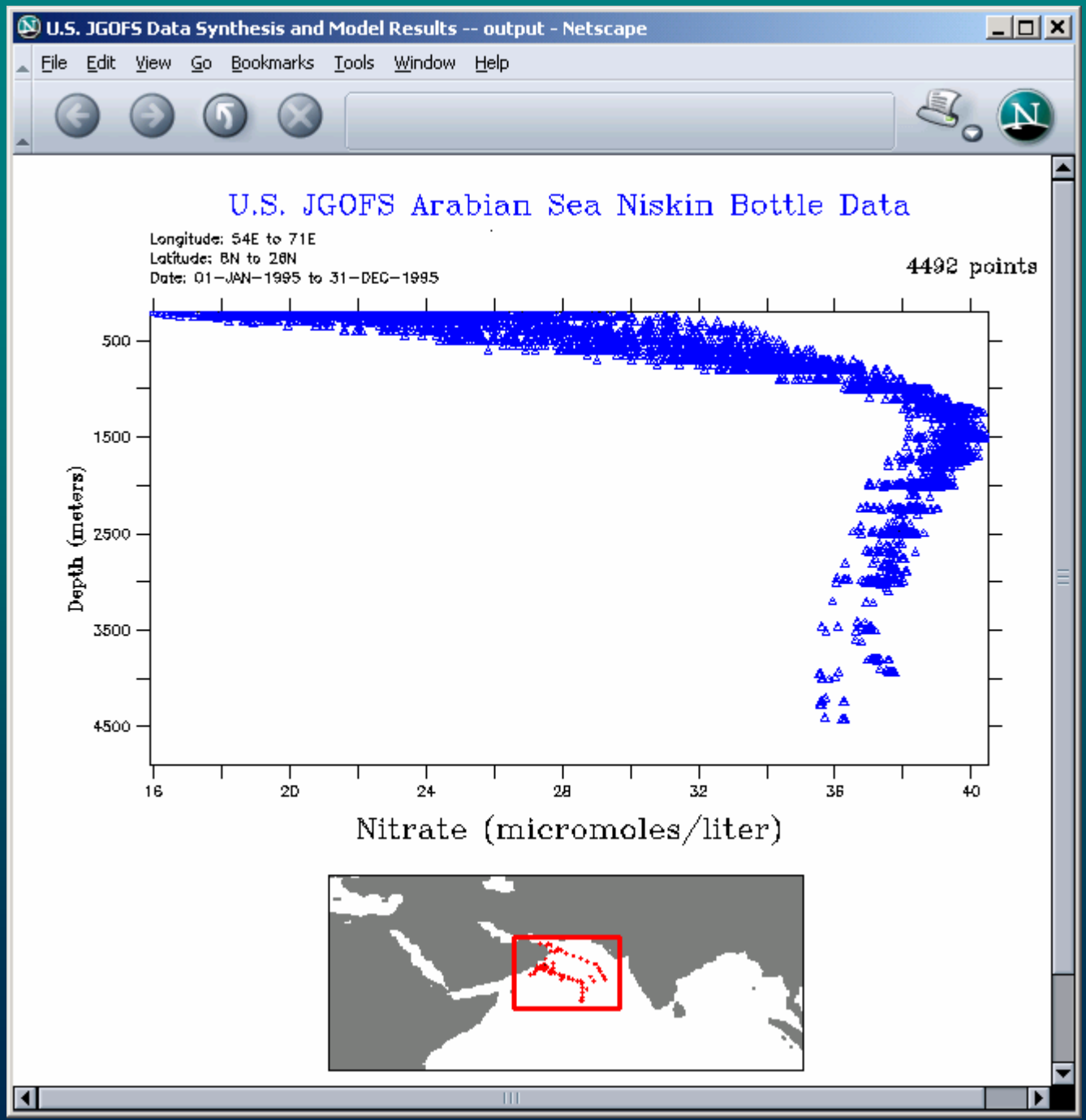
26 May 2005



LAS v6

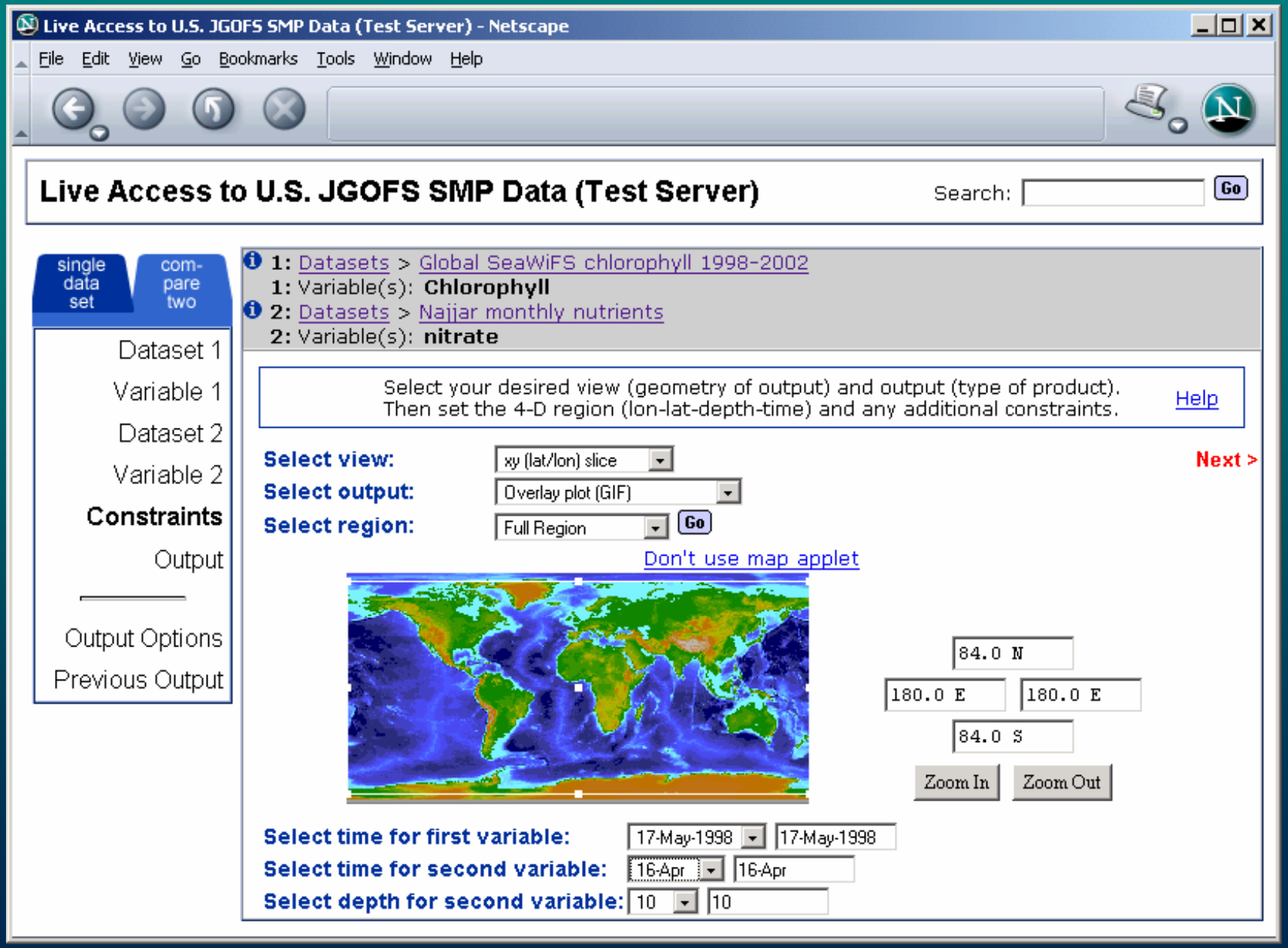
Property-Depth

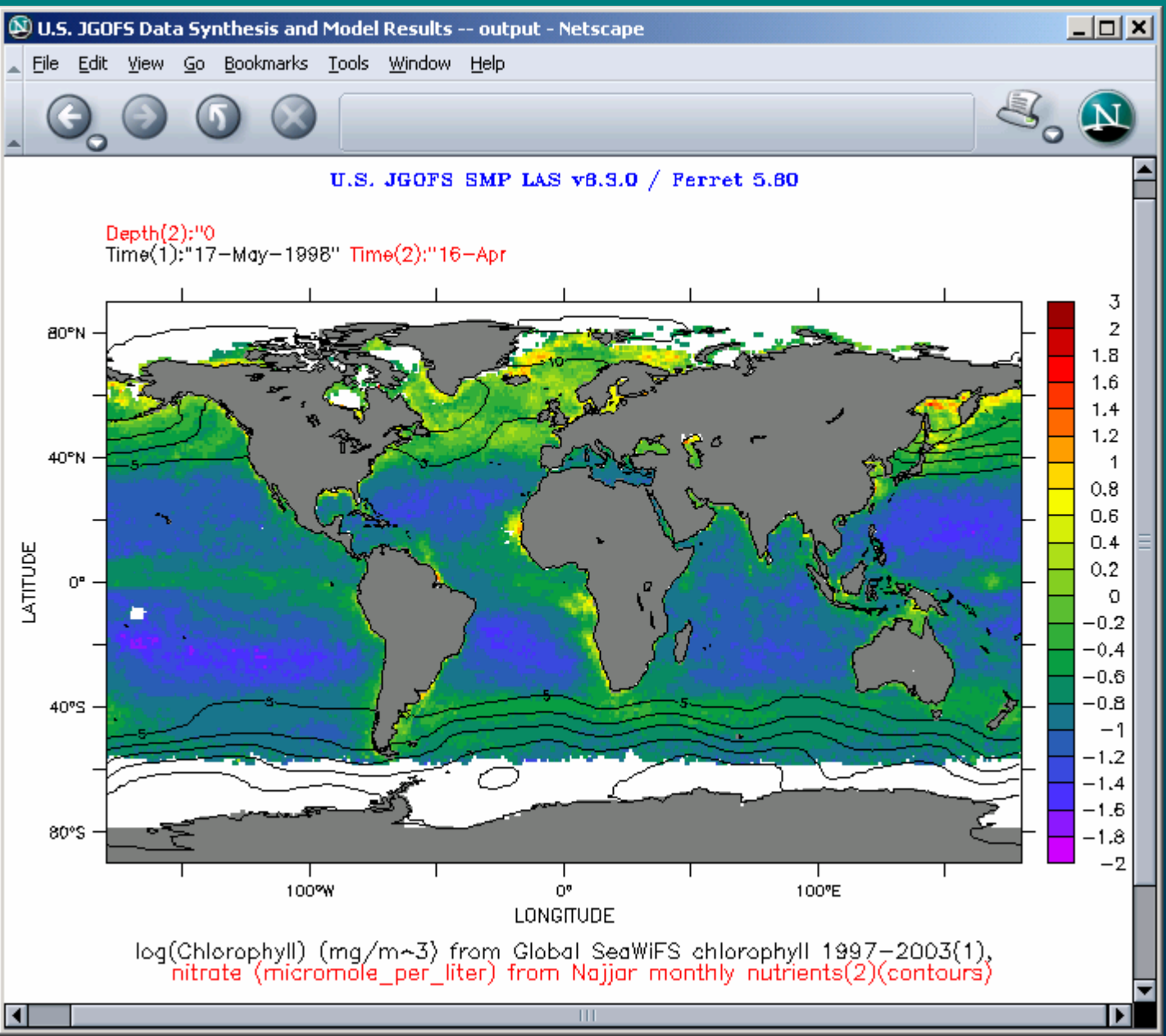
Arabian Sea Nitrate



from merged product generated by DMO from *in situ* Niskin bottle data

Comparison Overlay Plot





chlorophyll (May) shaded

SeaWiFS data contributed by: Yoder and Kennelly

surface nitrate (April) contours

Nutrient Climatologies contributed by: Ray Najjar

U.S. JGOFS Data System Summary

- supports a variety of data formats
- OP*e*NDAP used to access data collection
- coupled metadata and data
- supports data subselection
- offers variety of products for download

Lessons Learned . . .

- encourage data managers to collaborate
  - ✓ other data managers within program
    - ❖ JGOFS DMTT
  - ✓ data managers from other programs
    - ❖ GLOBEC, LTER
  - ✓ program investigators and participants
    - ❖ attend conferences and workshops
    - ❖ offer data system tutorials

Lessons Learned . . .

- publish data reports
  - ✓ archive data in one place
  - ✓ easy access to project results
  - ✓ most complete and accurate form of database

data reports published on CD-ROM

Lessons Learned . . .

- plan early for final archive of program results
  - ✓ digital records
  - ✓ don't forget the boxes of stuff !

Lessons Learned . . .

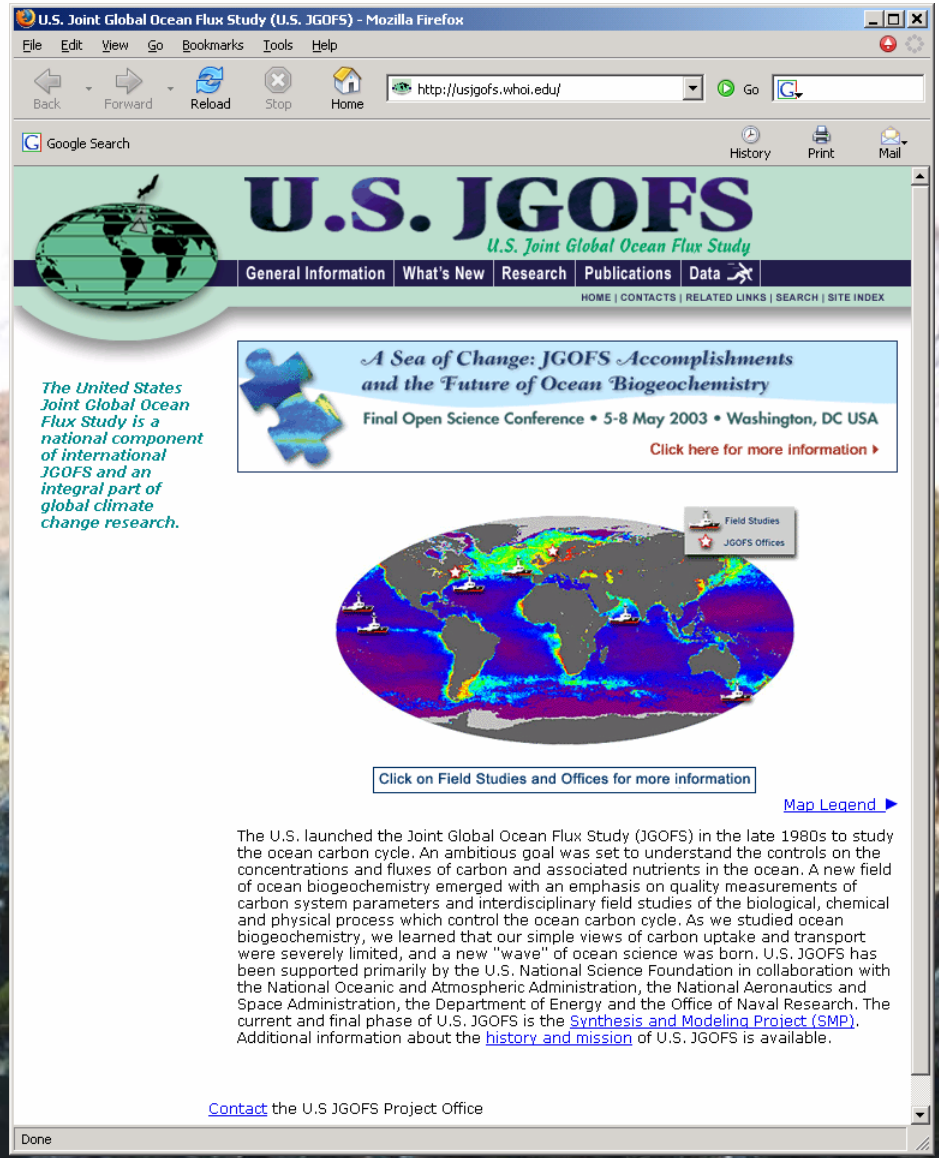
- develop a data policy, publicize and follow it
- establish protocols at program start with mechanism for adaptation when necessary
- facilitate contribution of data to collection
- metadata is of critical importance
  - accurate, complete, available with data
- quality assurance is an ongoing process
- begin synthesis early

Lessons Learned . . .

- provide timely, easy access to project results
- develop and maintain simple, reliable interface to data collection
- encourage data managers to collaborate
- publish data reports
- plan early for final archive of program results

Challenges

- functioning amid the chaos
  - maintaining a healthy data management system amid the chaos of rapidly changing information technology
  - distinguishing between enabling and disruptive technologies
- data and results – what to preserve?
  - raw and processed data, synthesized products, model code, inputs, results
- increasing volume and diversity
- long-term preservation of data



U.S. JGOFS web site
http://usjgofs.whoi.edu