A New Data Management System for Biological and Chemical Oceanography

Robert C. Groman[1], Cynthia L. Chandler[2], M. Dicky Allison[1], David M. Glover[2], Peter H. Wiebe[1]

[1]Biology Department, [2]Marine Chemistry and Geochemistry Department, Woods Hole Oceanographic Institution

Abstract

The Biological and Chemical Oceanography Data Management Office (BCO-DMO) was created to serve PIs principally funded by NSF to conduct marine chemical and ecological research. The new office is dedicated to providing open access to data and information developed in the course of scientific research on short and intermediate time-frames. The data management system developed in support of U.S. JGOFS and U.S. GLOBEC programs is being modified to support the larger scope of the BCO-DMO effort, which includes ultimately providing a way to exchange data with other data systems.

The open access system is based on a philosophy of data stewardship, support for existing and evolving metadata and data exchange standards, and use of public domain software.

In the new BCO-DMO data system, project and dataset metadata records designed to support re-use of the data are stored in a relational database (MySQL) and the data are made accessible by the JGOFS/GLOBEC object-oriented, relational, data management system. Data access is provided via any standard Web browser through a GIS application (Open Source, OGC-compliant MapServer), or a directory listing from the data holdings catalog. A custom search engine that facilitates data discovery is in development. In an effort to maximize data system interoperability, data will also be available via Web Services; and dataset descriptions will be generated to comply with a variety of metadata content standards. The office is located at the Woods Hole Oceanographic Institution and Web access is http://www.bco-dmo.org.

Introduction

The data management system, developed in support of the U.S. JGOFS and U.S. GLOBEC programs (Flierl et al., 1994, Groman and Wiebe, 1998, Glover et al., 2006), has served our past and current data management needs extremely well. It is a very flexible system designed to enable Web access to data via any standard browser and supports distributed data servers. However, in order to meet the needs of the new Biological and Chemical Oceanography Data Management Office (BCO-DMO) [see companion poster #OS41A-0164] we needed to enhance this system by providing a more uniform approach to storing and sharing metadata and by improving the Web user interface.

Metadata: Necessary Evil or the Basis of All Good Things

In order to fulfill the goal of the BCO-DMO Project, enabling data sharing and reuse of data collected by NSF supported biological and chemical oceanography projects, migration of our currently existing and all future metadata to a computer readable form was essential. Metadata are essential for understanding who collected the data, how the data were collected and processed, what data were collected and what the data fieldnames mean. Some would also say that where and when the data were collected are also important metadata. While this is true, we argue that these parameters are so essential, that they should be considered part of the datasets. In fact, it is arbitrary to distinguish between data and metadata. They are all data.

We spent a lot of time investigating available metadata and data format exchange standards (see for example the Marine Metadata Interoperability Project, http://www.marinemetadata.org/). We eventually realized that there is no uniformly accepted, single, standard for metadata and that we should be prepared to store and provide all the necessary metadata for data reuse in whatever format and organization fits our needs. Once this is done, we will be able to output metadata fields in whatever metadata standard people prefer. We chose to use a relational database model, implemented in MySQL, to store our metadata. This approach also enables changes and additions to the database tables as needed due to the inherit flexibility of the Structured Query Language (SQL) and MySQL.

We are building into the metadata database as much flexibility as possible. Our goal is to be able to accept data from all the data contributors involved in the supported NSF programs without requirements that impose undue hardship on them. The goal is to be able to display, synthesize, and share data when the data were collected by many different means, using many different instruments and, indeed, may even be called different names by different people.

People consider metadata solutions "a Sisyphean torture" (Shankaranarayanan and Even, 2006). However, it is the availability, accessibility, and completeness of metadata that enables people to share their data with others.

Metadata Database Schema

Location table: location_id, platform_id, latitude, longitude, depth_w, height, timezone, date_time

Platform table: platform_id, platform_name, platform_type (ENUM: mooring, vessel, ROV, …), deployment (aka cruiseid; no duplicates allowed), synonyms, chief_scientist_name_id, co-chief_scientist_name_id, start_date, end_date, location (text) [use separate table instead?]

People table: people_id, first_name, middle_name, last_name, title [ENUM: Dr, Ms, …], affiliation_id, address_line1, address_line2, city, state, country, zipcode, phone, fax, email

Affiliation table: affiliation_id, affiliation_name, acronym, subname

Funding table: funding_id, funder_name, acronym

Knowing who collected, contributed, and owns the data is critical. (purple dots)

Dataset_platform table: dataset_platform_id, dataset_id, platform_id, platform_activity [ENUM: deploy, recovery, turnaround], acquisition_description [data acquisition and methodology description], processing_description [data processing description], dataset_platform_url, version

Dataset_platform_people table: dataset_platform_people_id, dataset_platform_id, people_id, title (ENUM: student, PI, etc.)

Dataset_funding table: dataset_funding_id, dataset_id, funding_id, project_number

Program table: program_id, program_name, acronym, description, program_url, lead_pi_id, co_pi_id, start_date, end_date, logo_url, geolocations

Parameters table: parameters_id, standard_name, short_name, common_name, official_name, equivalent_name, units, fill_value, description

Enable crosswalk between differently named data by using controlled vocabularies to support interoperability (green dots)

Project table: project_id, dataset_id, program_id, project_name, acronym, description, project_url, lead_pi_id, co_pi_id, contact_id, start_date, end_date, logo_url, geolocations

Start_stop Table: start_stop_id, dataset_platform_id, platform_id, start_date, end_date, latitude (or Null), longitude (or Null), station – see note 9, tow – see note 9, cast – see note 9

Dataset_platform_status: dataset_platform_status_id, dataset_platform_id, status, comment

Dataset_status: dataset_status_id, dataset_id, entered_by_id, next_action_date, status {ENUM}, comment

Dataset table: dataset_id, dataset_name, project_id, dataset_url, dataset_description, validated (flag), version

Dataset_parameters table: dataset_parameter_id, dataset_id, parameters_id, supplied_name, supplied_units, conversion_utility, fill_value, description

Modification_history table: modification_history_id, dataset_id, entered_by_id, description [includes table name, local id, and action]

Dataset_instrument table: dataset_instrument_id, dataset_id, instrument_id, instrument_description

Instrument table: instrument_id, instrument_name, acronym, description

How the data were collected must be preserved as well (pink dots)

October 11, 2007
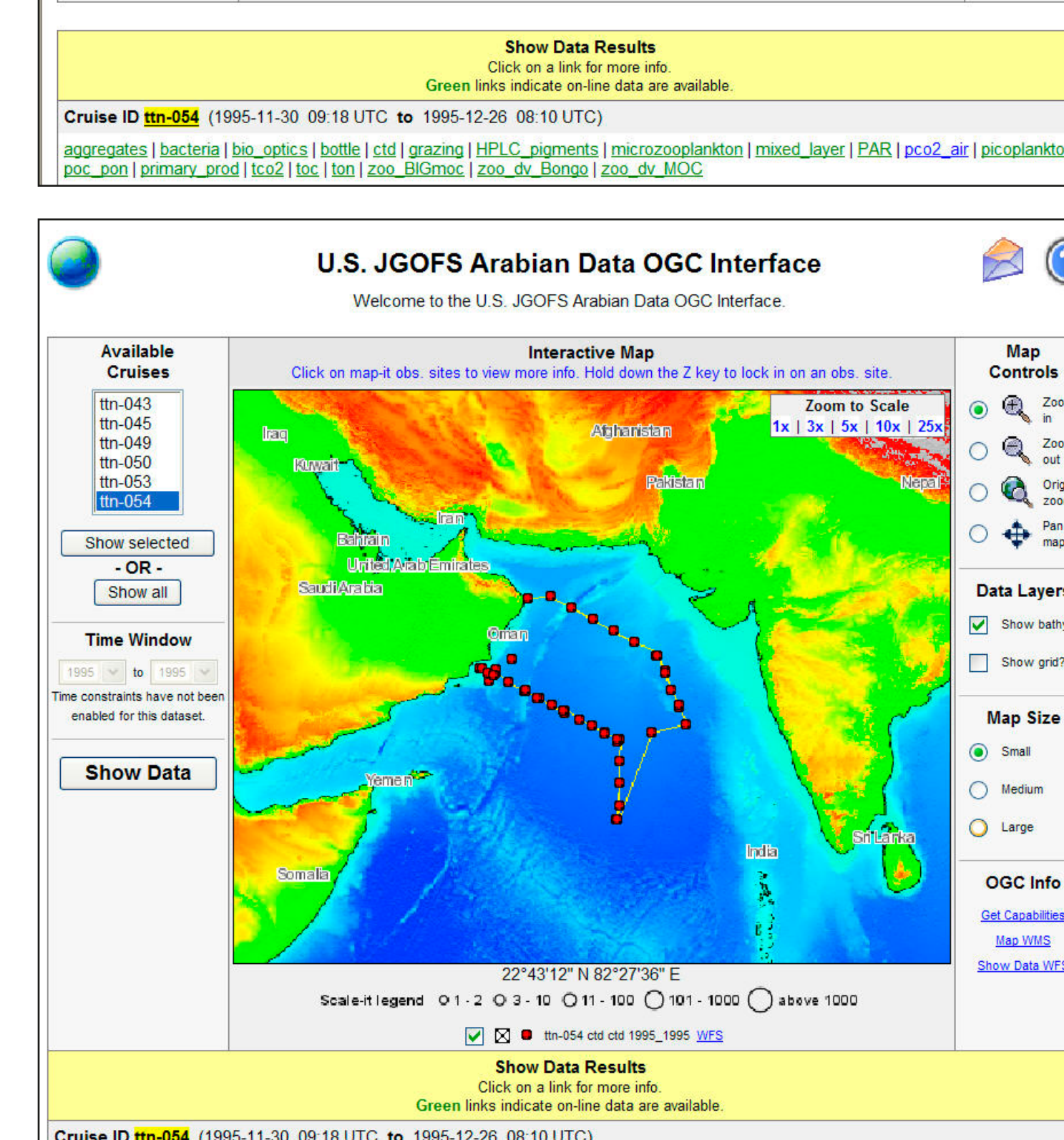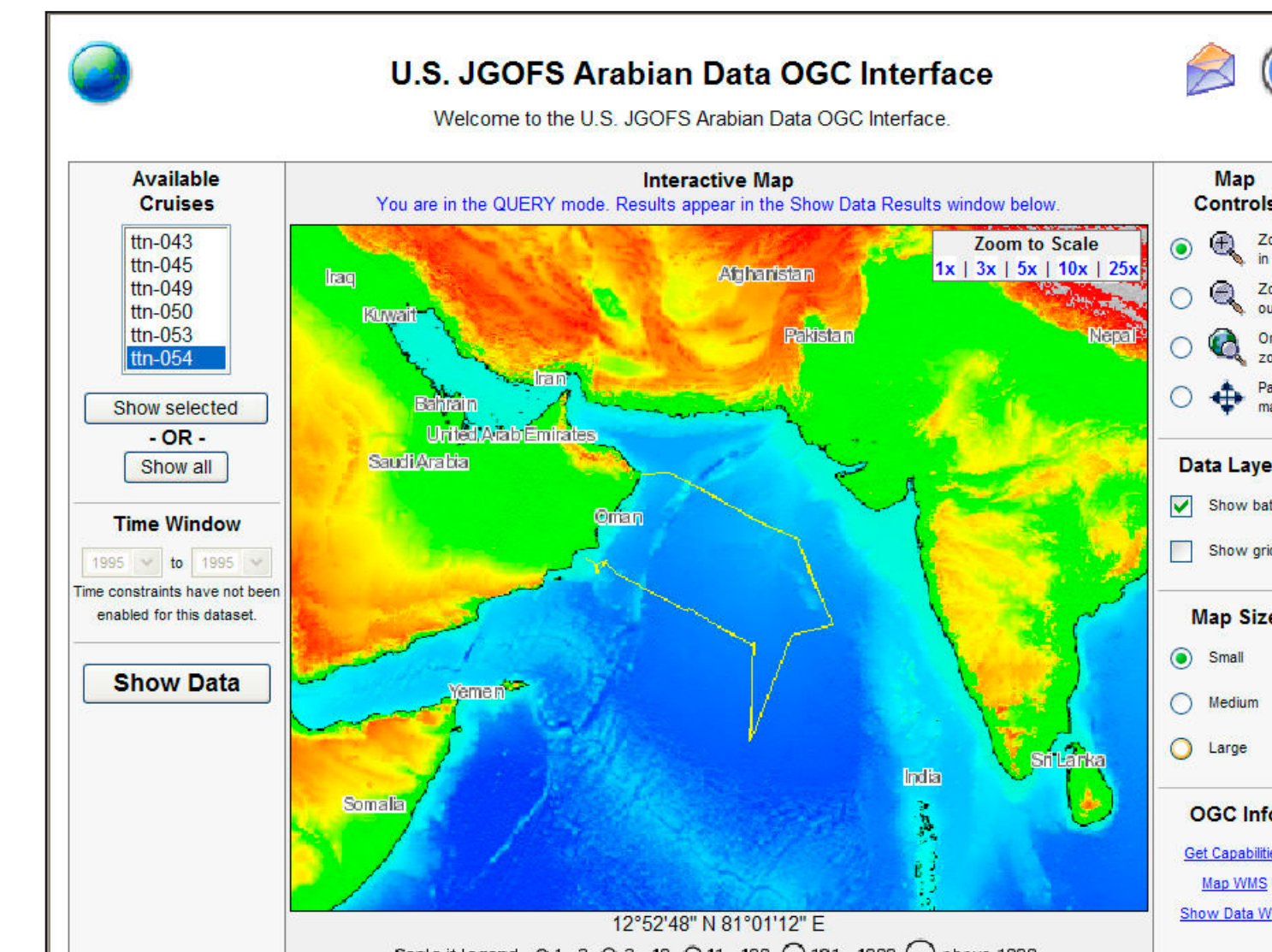
Visualization and Data Interoperability

- MapServer (see MapServer, 2007) interface with OGC WMS and WFS access.
- SQL/OBIS (DiGIR) interface to related biological, chemical, and physical oceanographic data will be supported.
- Google Earth KML output format will also be supported.
- Data can be downloaded in several user-selectable formats: comma or tab-separated ASCII, Matlab binary format or NetCDF files. Other formats (e.g. Ocean Data View) will be added in the future.
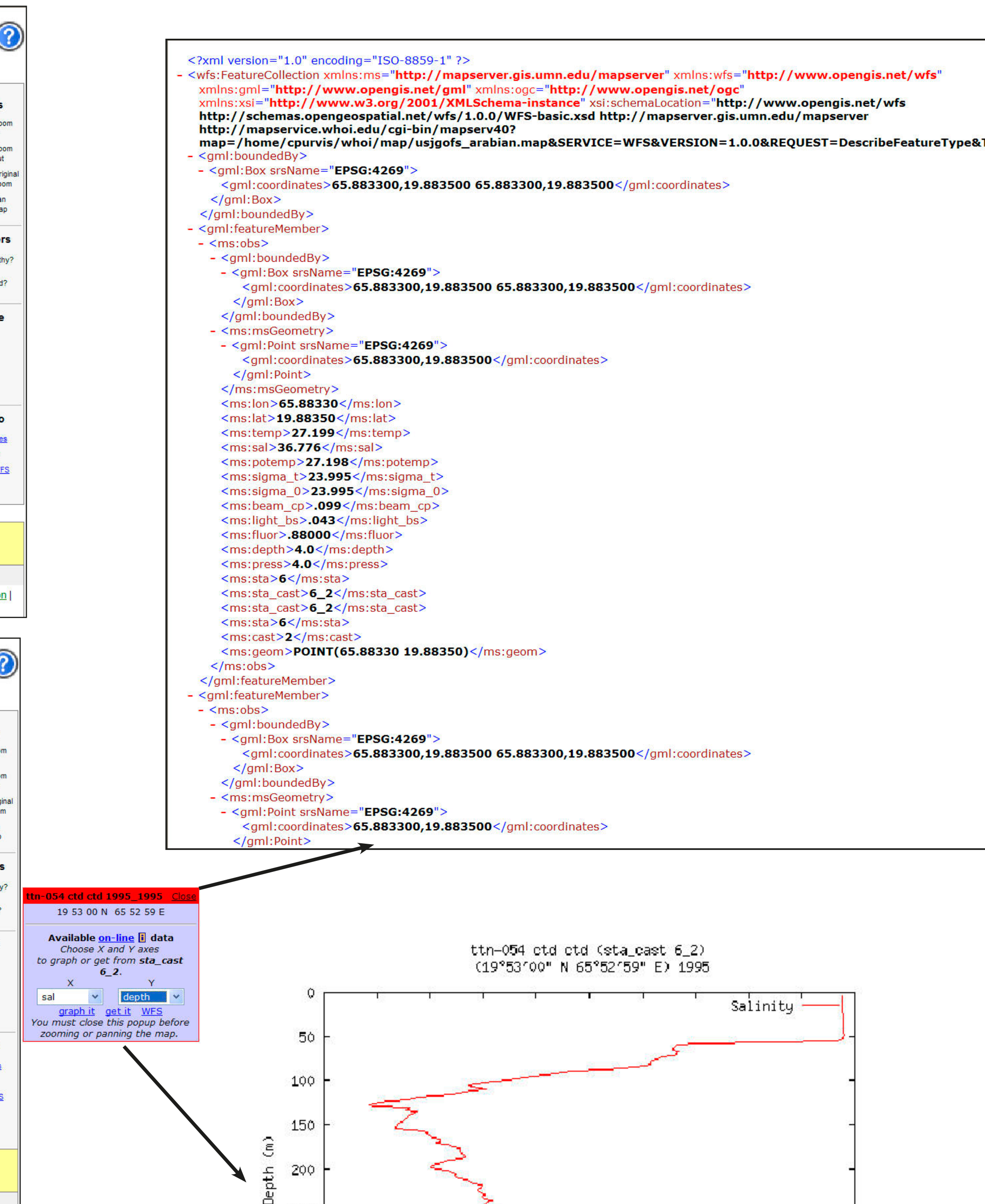
As illustrated by the two separate map images below, the U.S. JGOFS and U.S. GLOBEC data sets are currently accessed through separate but similar MapServer interfaces. Our goal is a single, integrated map display, supporting all available data sets, retrievable by such attributes as geographic location, date and time, funding sources, project/program names, and names of scientific investigators.

U.S. JGOFS Data MapServer Interface
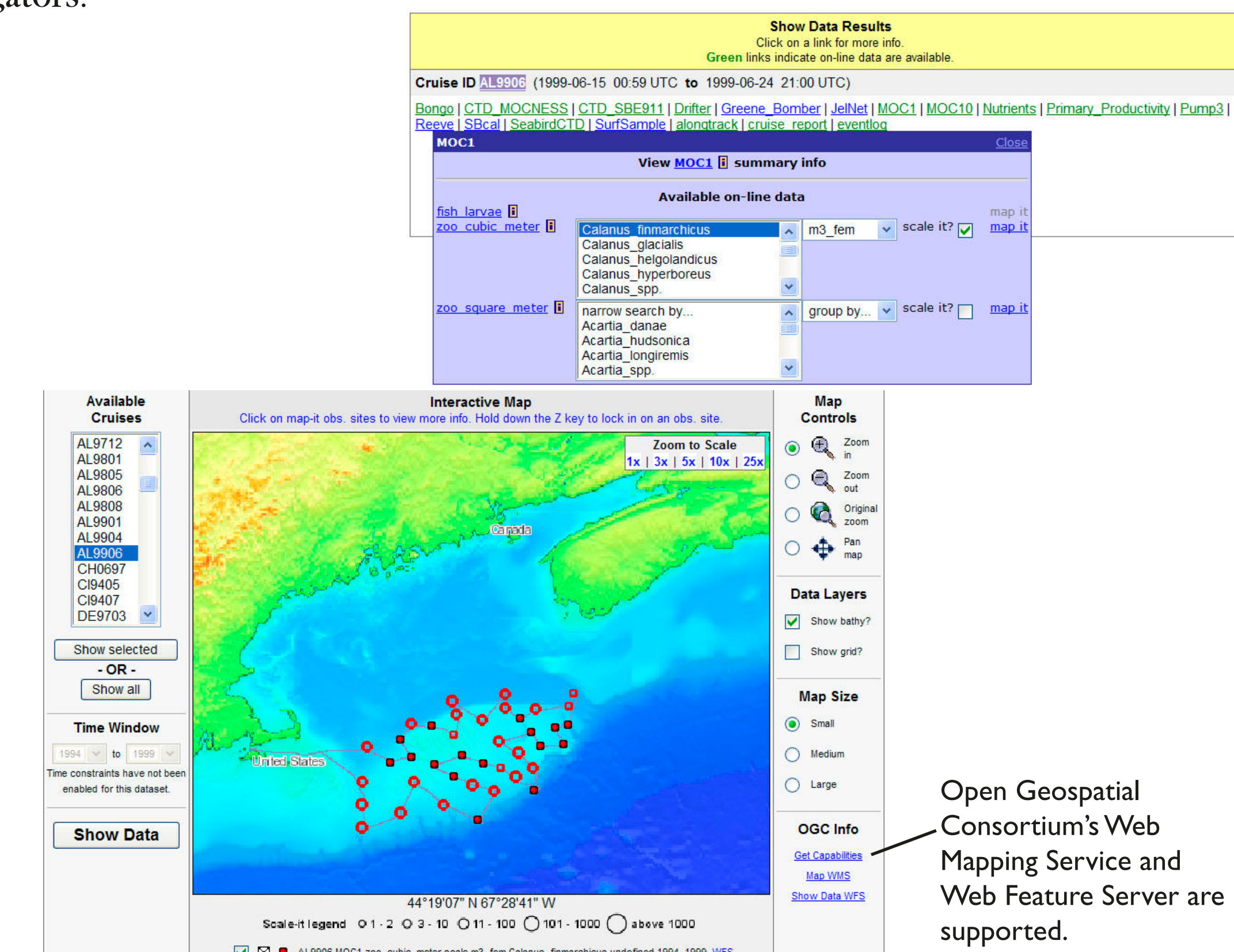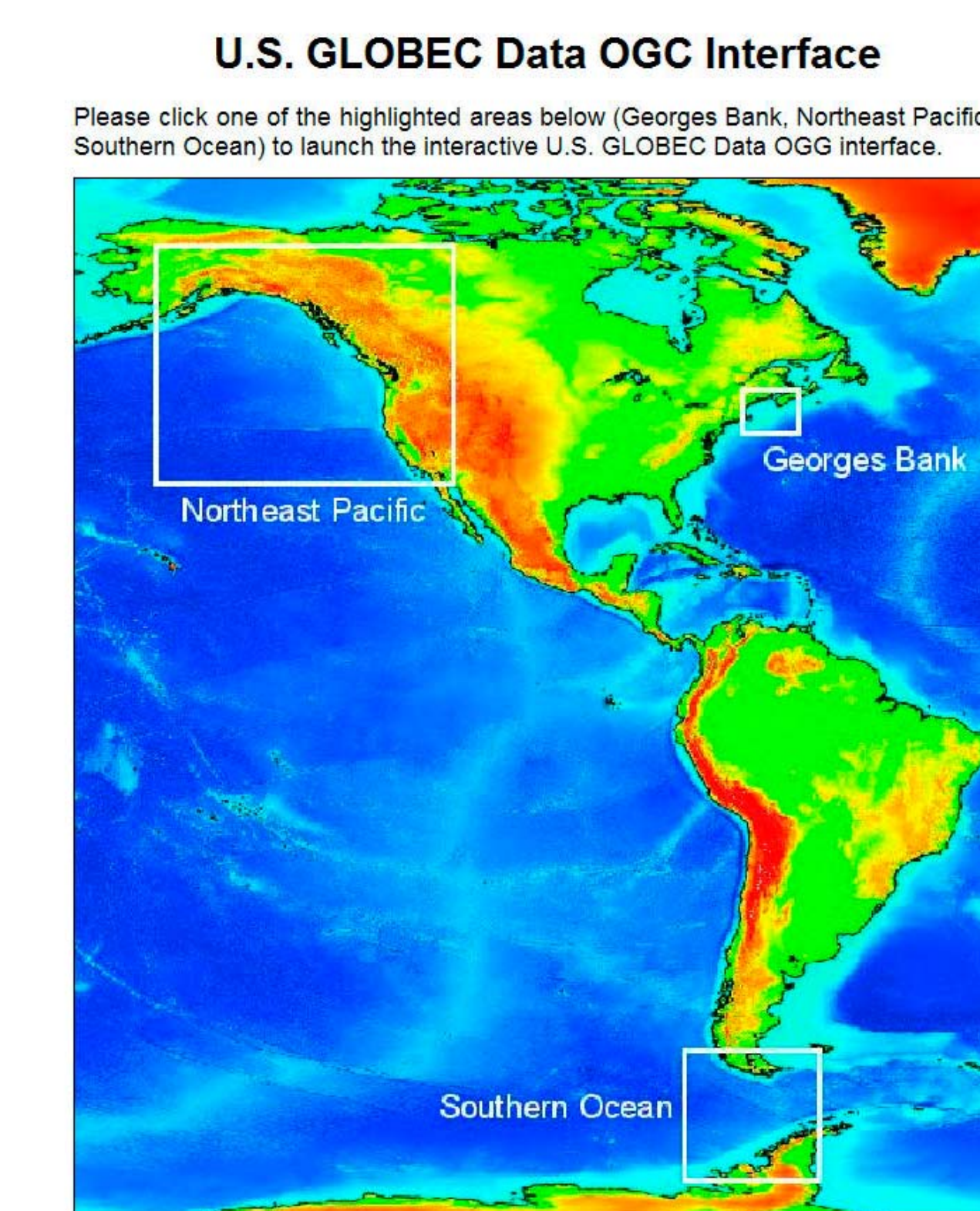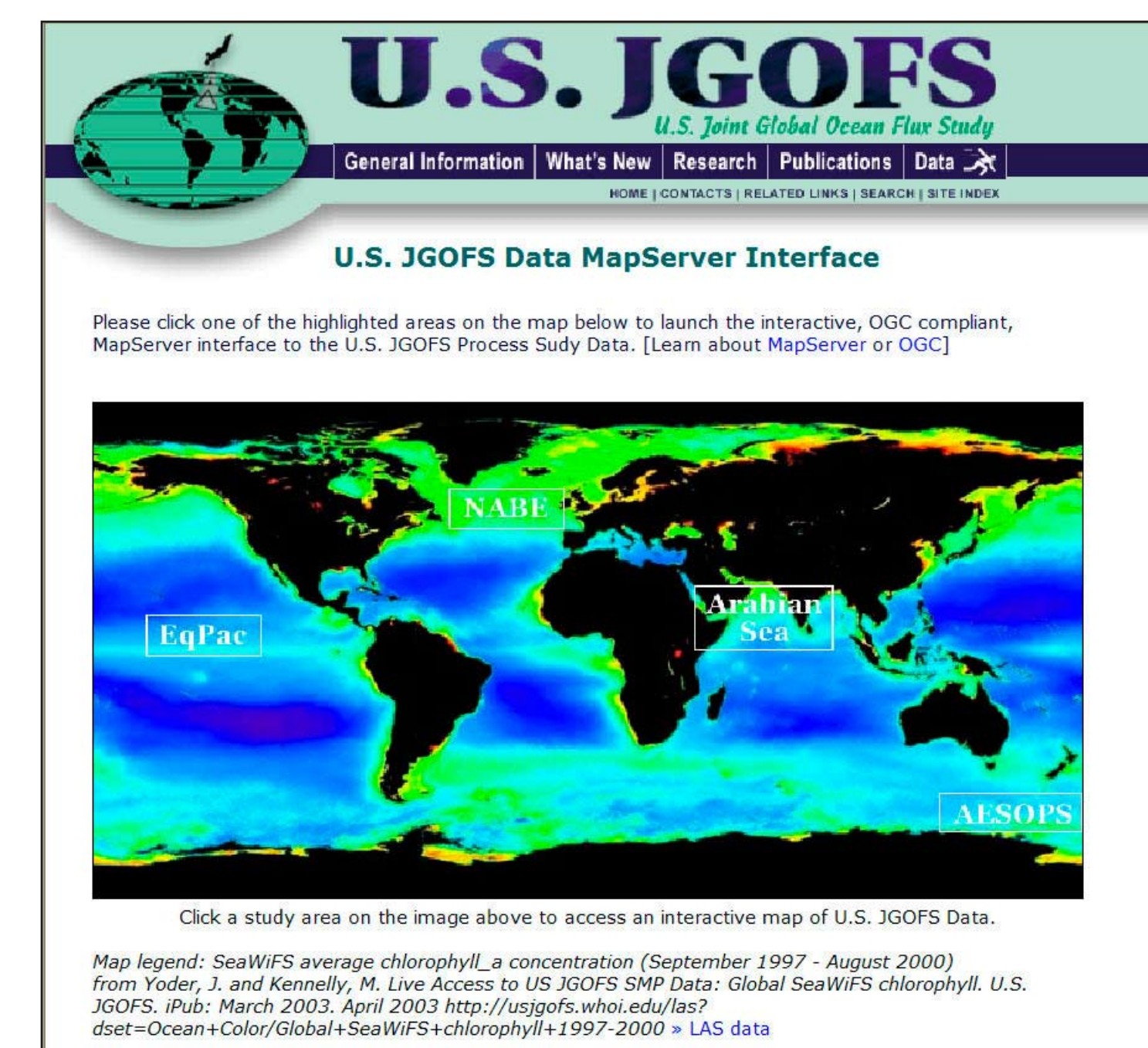
After clicking on an area of interest in the map (above), one sees all the available cruise tracks and can then highlight one or more cruise names and click on "Show Selected".

U.S. GLOBEC Data OGC Interface

Open Geospatial Consortium's Web Mapping Service and Web Feature Server are supported.
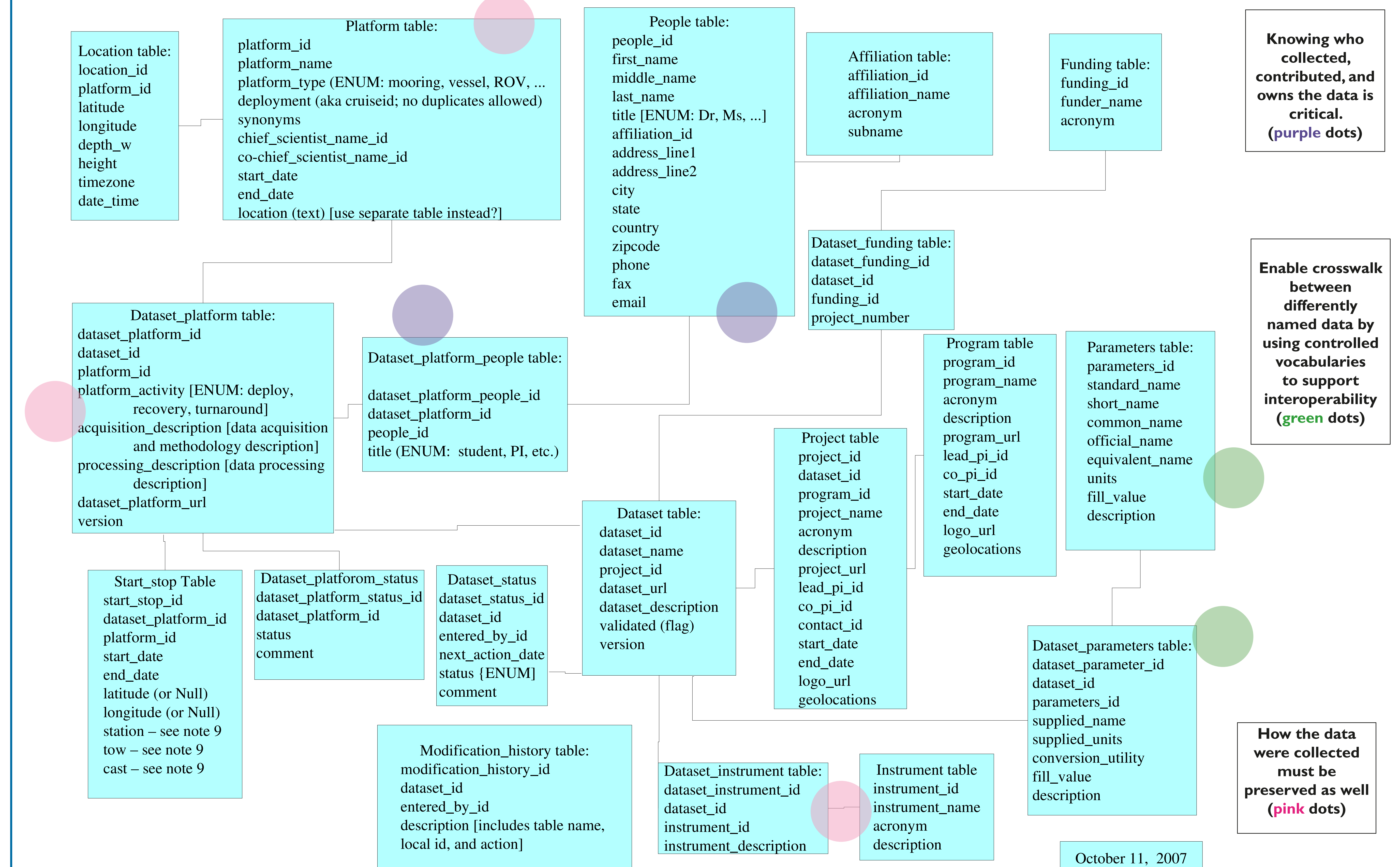
Clicking on a data set (e.g. ctd as shown above) displays a box of available data. Clicking on the "map it" link displays where the data are located, as colored circles. Cicking on one of the circles displays a small box with the station information and the ability to make an X-Y graph, Display the data in tabular form, and download the data via the Web Feature Service protocol.

Conclusion

The fundamental design and implementation features of our metadata database and data management system to support the BCO-DMO Project include the following:

- Keep it simple
- Keep in complete
- Support the MapServer application
- Support data interoperability
- Allow distributed data clients and data servers
- Recognize lack of comprehensive standards

References

Flierl, G.R., J.K.B. Bishop, D.M. Glover, and S. Saranjpe. JGOFS Data System Overview, On-line report, http://globec.whoi.edu/globec-dir/globec-data-system-description.html, 1994.

Glover, D.M., C.L. Chan dler, S.C. Doney, K.O. Buesseler, G. Heimerdinger, J.K.B. Bishop, and G.R. Flierl, 2006. The U.S. JGOFS Data Management Experience, Deep-Sea Res., II/, *53"(5-7), 793-802.

Groman, R.C. and P.H.Wiebe, Data Management in the U.S. GLOBEC Georges BankProgram, Marine Technology Society's Ocean Community Conference 1998 proceedings, November 1998.

MapServer, University of Minnesota, http://mapserver.gis.umn.edu/, 2007

Shankaranarayanan, G. and A. Even. The Metadata Enigma, Communications of the ACM, Vol. 49, No. 2, pp 88-94, February 2006.

Acknowledgments