Interoperability Enhancements to the JGOFS/GLOBEC Data Management System

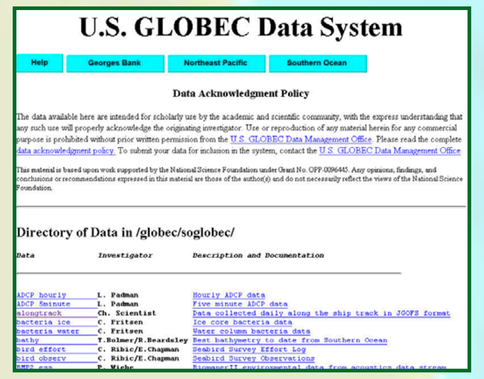
R.C. Groman[1], P.H. Wiebe[1], M.D. Allison[1], G.F. Heimerdinger[1], A. Schloss[2], W.J. Armstrong[2], W.S. Little[3], N.H. Vine[4]

[1]Woods Hole Oceanographic Institution, Woods Hole, MA 02543 • [2]University of New Hampshire, Durham, NH 03824
[3]Integrated Statistics, Woods Hole, MA 02543 • [4]Consultant, Woods Hole, MA 02543

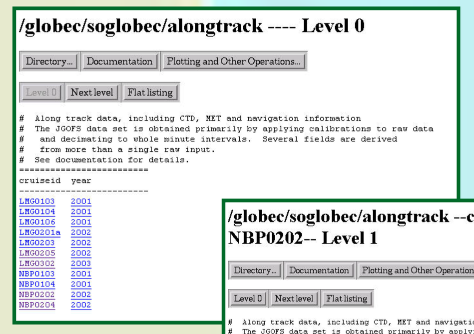
Abstract

The U.S. GLOBEC Program is a large, multidisciplinary effort, involving over 200 scientific investigators from over 70 research, educational, and governmental organizations. There are three program modules in the U.S. GLOBEC science effort: Georges Bank, Northeast Pacific, and Southern Ocean. Since each of these programs is in a different study phase, they require different kinds of data management support. The backbone of the support is provided by the JGOFS/GLOBEC Data Management software package, which enables the user community to access data sets via the Internet using standard web browsers. This package is used by several other ongoing data intensive projects including FleetLink, Martha's Vineyard Coastal Observatory, Northeast Consortium, and the recently completed U.S. JGOFS projects. But as more data become available from different data sources the need for data management interoperability among different data management systems is receiving greater attention. Interoperability issues include data assurance, common scientific field name use, metadata preparation, data discovery, and data accessibility. The capabilities and features of the JGOFS/GLOBEC system are being enhanced to address these and other interoperability issues so that other sites may access data served by the system. For example, an Open Geospatial Consortium based interface to our data is being incorporated to enable an alternate way to browse and download data served by the system. In addition, progress is being made in interfacing with the University of New Hampshire's GeoZui3D 3D visualization program.

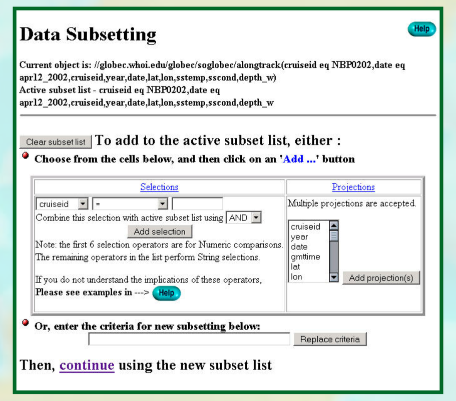
Data In

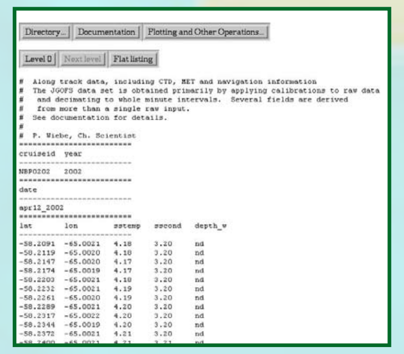


The main data server controls what data are made publicly available through the use of the data directory. This directory provides the clickable links to the actual data and on-line documentation. One does not need to know where these data reside in order to access them.



The data system supports data selection and projection. The selection option allows one to select data that match your criteria, such as the data where the temperature is greater than 10 degrees. The projection option allows one to select only those data fields of interest, i.e. perhaps only temperature, but not salinity.





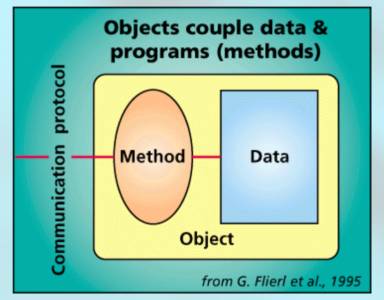
Projection (which variables?)

Selection (which cruise, which day, which location?)

Introduction

The JGOFS/GLOBEC data management system (Flierl, 1992; Groman and Wiebe 1997) was initially developed in the late 1980's to provide on-line access to data collected by the U.S. Joint Global Ocean Flux Study (U.S. JGOFS) program. In the early 1990's the networking protocol was modified to take advantage of the newly standardized Hypertext Transport Protocol (http) protocol used by the ever-growing World Wide Web and client browsers. The system supports distributed data servers and distributed clients. The attributes that make this software appealing include:

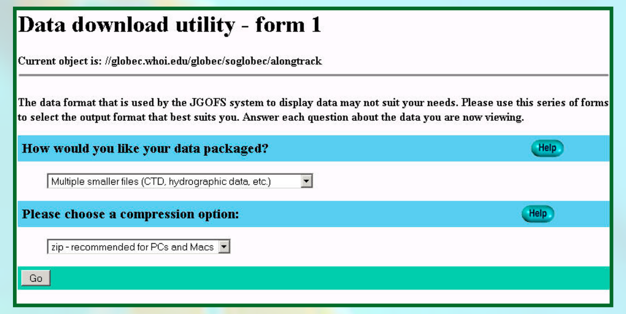
- Simple installation on Unix platforms,
- Minimal effort to get existing ASCII, tabular data on-line,
- Data are served directly by the data contributor, rather than copied to a "central data server." This avoids the problem of the served data being out of date. Also, data listings and data displays are created at the time of the request to insure that these data and images are the most up to date available. We refer to this feature as "data driven".
- Object oriented structure so that each data set and its associated serving software are tied together into a "data object." It enables the system to manipulate data objects independently of how the original data are actually obtained from the data file(s). Indeed, the data may not come from a file at all, but could be the output of a modeling application.
- Automatic restructuring of the data so that data can be effectively served as if they were stored in a relational database.
- In addition to ASCII data, images and movies are also readily made available on-line.
- Access to the data is through any standard web browser.

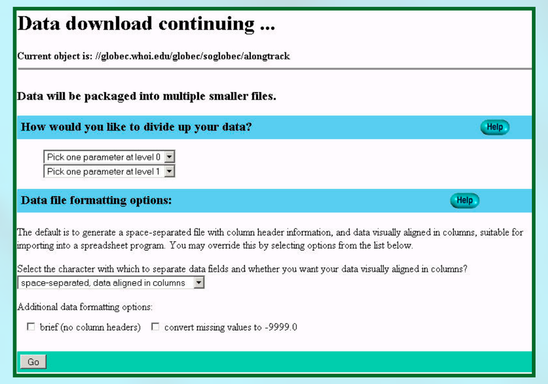


from G. Flierl et al., 1995

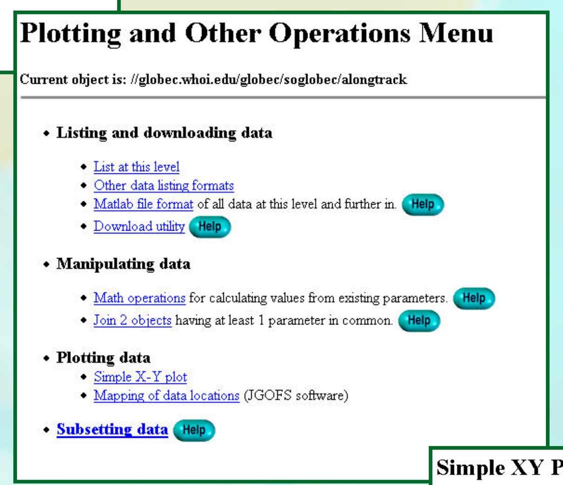
The original design criteria for the system assumed that investigators, managers, teachers and students would view the available list of data objects via the "Directory Page" and select a data object of interest. Once selected, it is easy to create simple X-Y plots of these data to determine whether these are the data of interest and to identify major attributes and features of these data. At this point, the original developers assumed one would download these data for further analysis and plotting using one's own analysis and plotting tools. Initially, the download option was accomplished by a simple "Save as" option from the file menu, but over time, additional download options have been added to accommodate the needs of the investigators. These include ASCII files, organized as a single flat file or as multiple files organized according to the data collected, MATLAB© binary file, and comma separated or tab separated tables. We also are working on a NetCDF output option.

Data Out



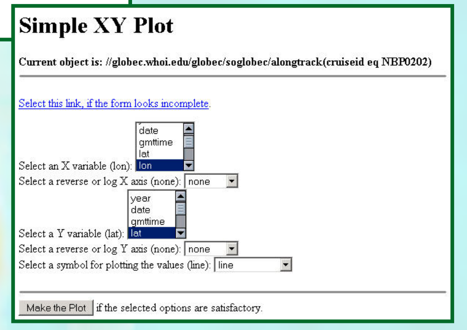


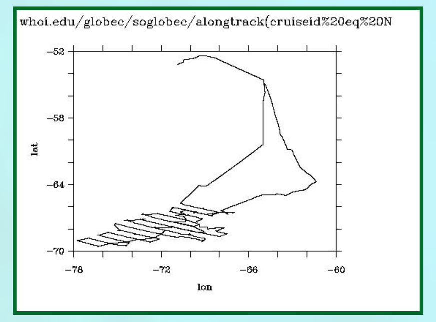
Download



There are several download options, including ASCII flat file, binary MATLAB file format, and in the case of one or more ASCII files, reorganized according to your needs. In addition, a simple x-y plotting option exists to help determine if the data are of interest. These operations are always "data driven"; that is, they are specific to the data object selected.

Quick Look

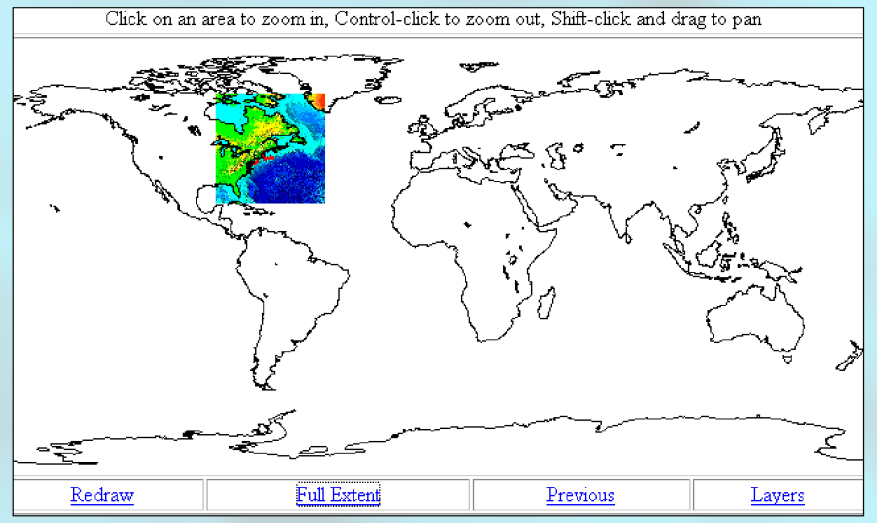




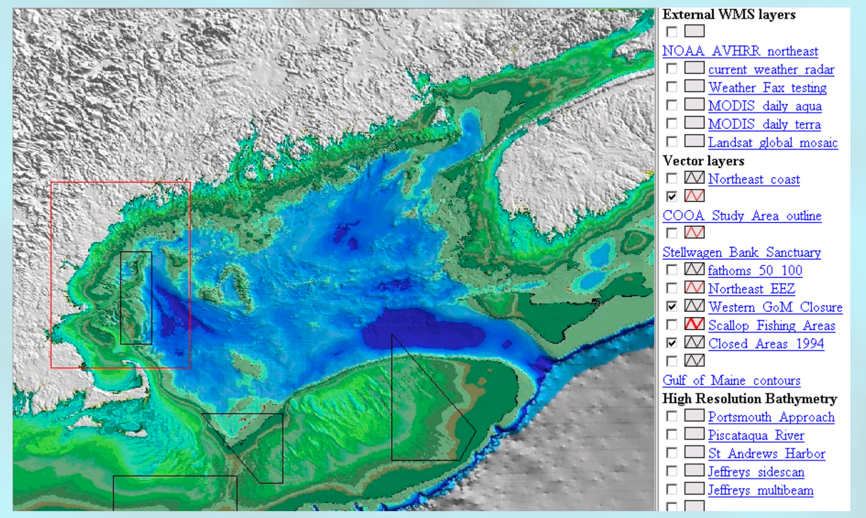
Data Discovery

Visualizing Available Data

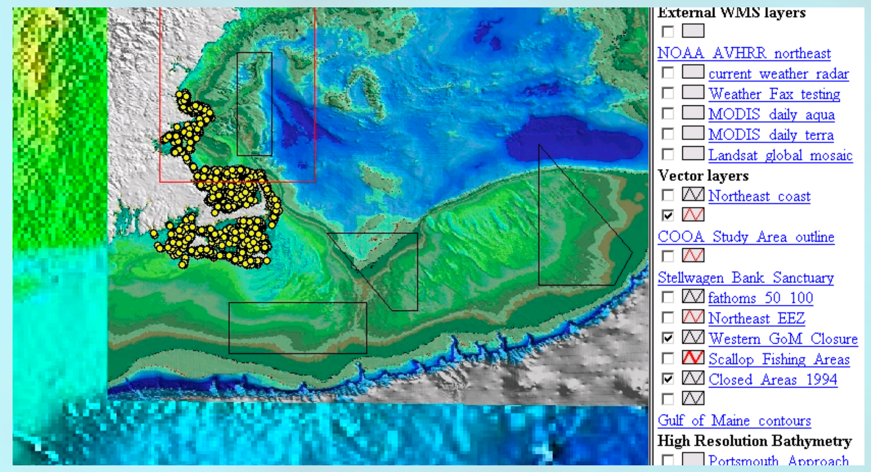
The expectations of web users and the technology have evolved much since the mid 1990's and many now expect to see clickable maps of available data with live links to these data. Indeed, the U.S. JGOFS program decided to implement a Live Access Server in order to create a clickable image of their available data. The Live Access Server (LAS) is a very flexible Web server designed to provide flexible access to geo-referenced scientific data (PMEL, 2006). LAS was not originally designed to work with non-gridded data sets. But programmers in Seattle and Woods Hole put together the additional modules necessary to allow LAS to work with U.S. JGOFS process study profile data. The customized LAS interface allowed investigators to navigate and visualize the merged data products in a much more convenient and rapid manner (Chandler, 2006). In this case, a relatively small number of merged data objects (data sets), about a dozen, were prepared and input to LAS. Our approach is similar, except we will be using the Open Geospatial Consortium Web Mapping Service and Web Feature Service standards to visualize the available data and ultimately provide clickable links to the underlying data. Another difference between the U.S. JGOFS effort and this effort is that we intend to provide access to each individual data object, and to the extent possible, retain the "data driven" nature of the JGOFS/GLOBEC data management system.



Simple OGC WMS (Web Mapping Service) Client

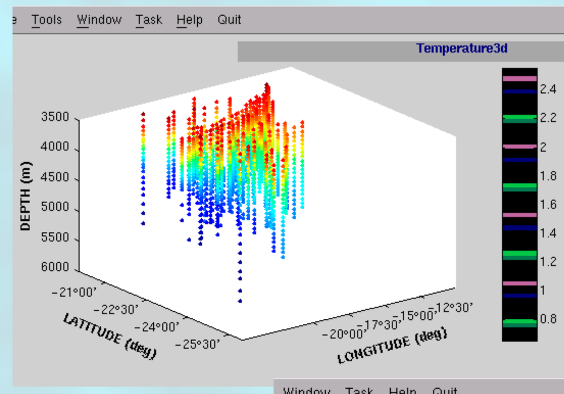


Gulf of Maine high-resolution bathymetry with fisheries closed areas marked in red and black.

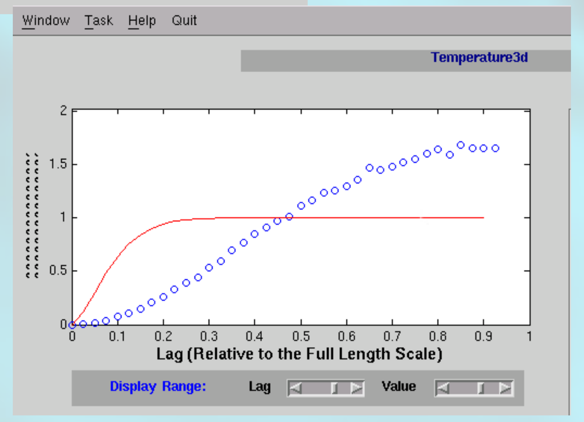


Above plot with cod data added as yellow circles.

Gridding

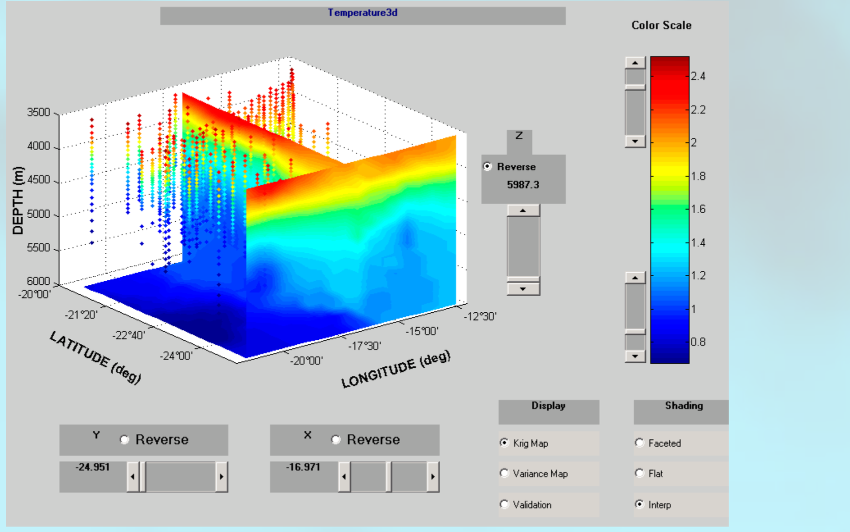


Temperature data loaded



Temperature data variogram

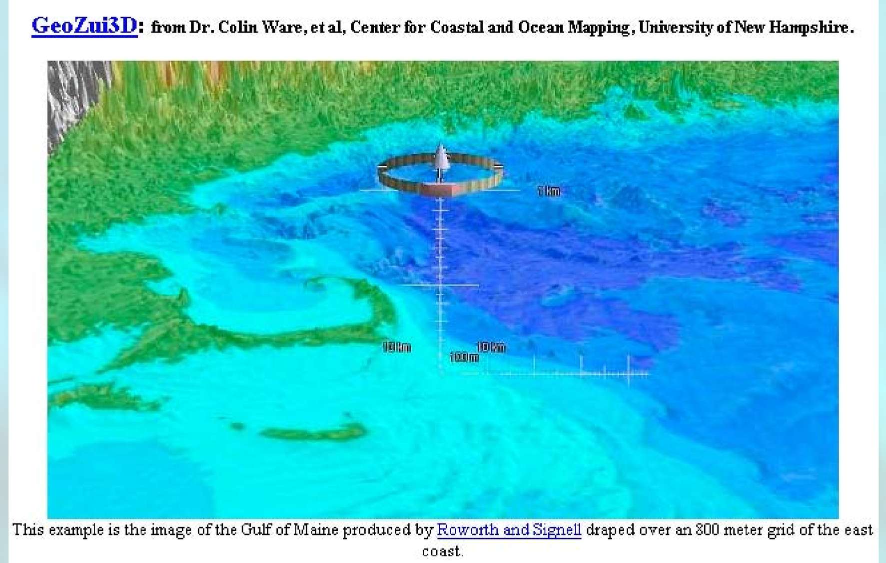
To prepare an irregularly spaced data set into a gridded data set, several tools can be used. EasyKrig, a Matlab based, 2D/3D gridding tool was written by Dezhang Chu for the U.S. GLOBEC program to objectively map data and create gridded data from experimental measurements. It is available for download from ftp://globec.whoi.edu/pub/software/kriging/easy_krig. A Web based objective mapping option (also based on EasyKrig) is planned so that a researcher can see an objectively mapped display. This capability has to be used with due caution, however, because of the inherent problems of generating artifacts in the resulting, gridded data set. Once data are objectively mapped, they can be imported to an analysis and display tool like Matlab or to a 3D viewing program like GeoZui3D for further investigation.

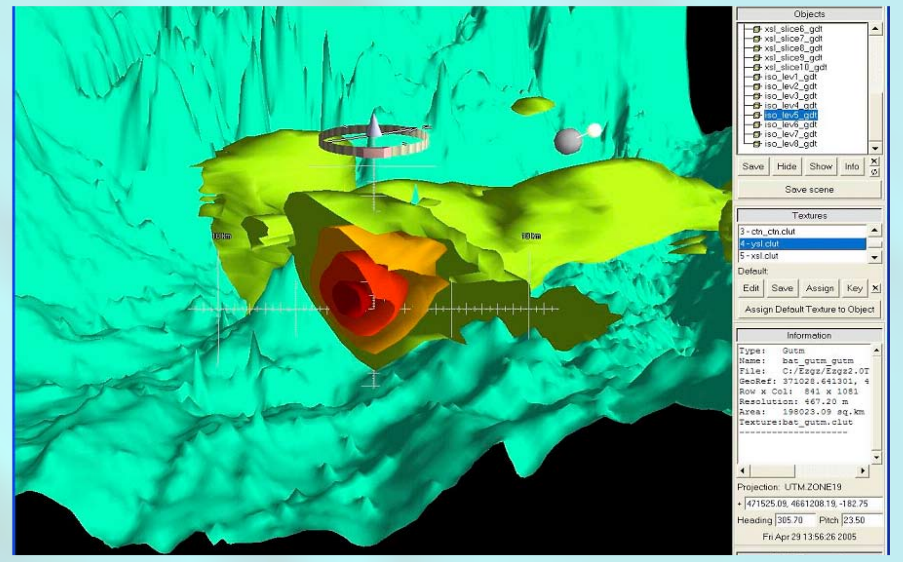


Temperature data final plot

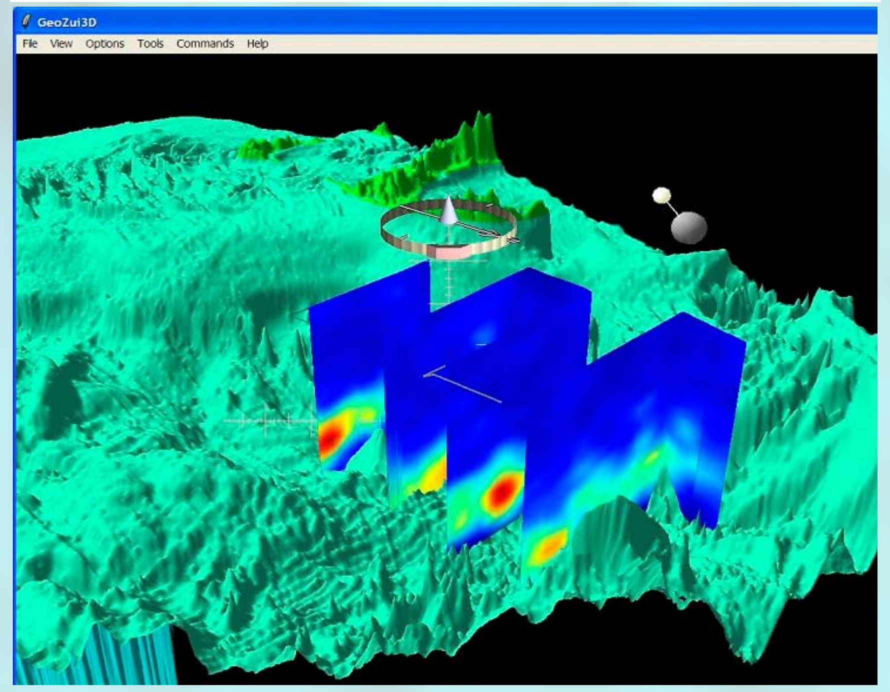
3D Visualization

While most if not all scientific investigators have their own set of analysis and visualization tools, it is recognized that others might not have these tools available. Because of the nature of the data, existing in three dimensional space and time, it was apparent that a 3D visualization tool would be most useful. Our initial efforts have focused on using the University of New Hampshire's GeoZui3D program. The GeoZui3D program is a highly interactive, PC Windows-based, 3D visualization system that supports a number of different research projects in the Center for Coastal and Ocean Mapping at the University of New Hampshire (UNH). It is a Zooming User Interface - hence 'Zui'. It is geo-referenced—hence GeoZui. It emphasizes interactive 3D solutions - hence GeoZui3D. This public domain program is available from UNH at http://www.ccom.unh.edu/vislab/GeoZui3D.html. Tools are being developed to enable easier preparation of data, extracted from the JGOFS/GLOBEC data management system, for use within the GeoZui3D program. A preliminary version of such a tool, called EZGZ (pronounced Easy-Geezy), has been used by several U.S. GLOBEC investigators to help improve its user interface and add the necessary features.

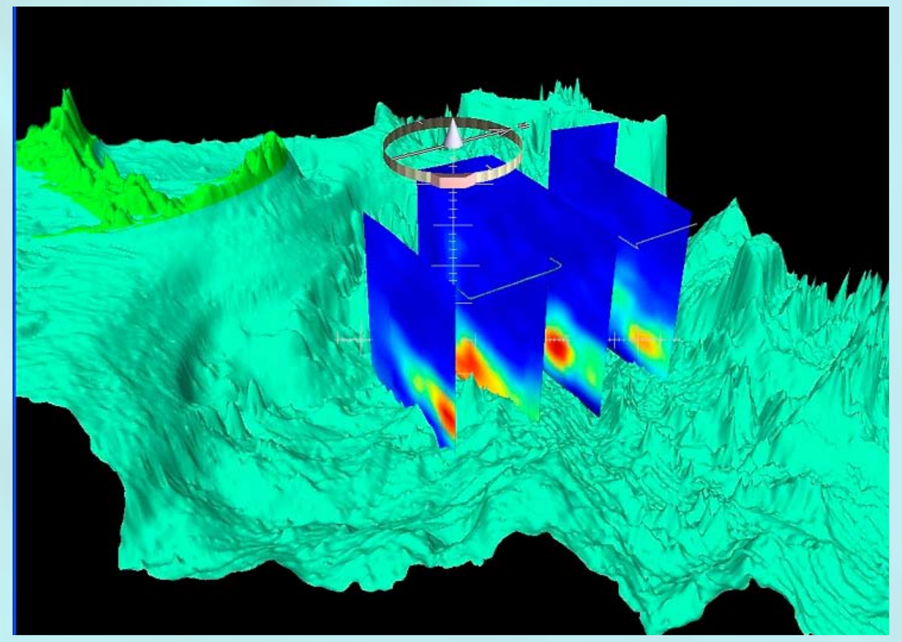




Calanus finmarchicus abundance isosurfaces.



Plot of gridded Stellwagen Bank Calanus finmarchicus abundance.



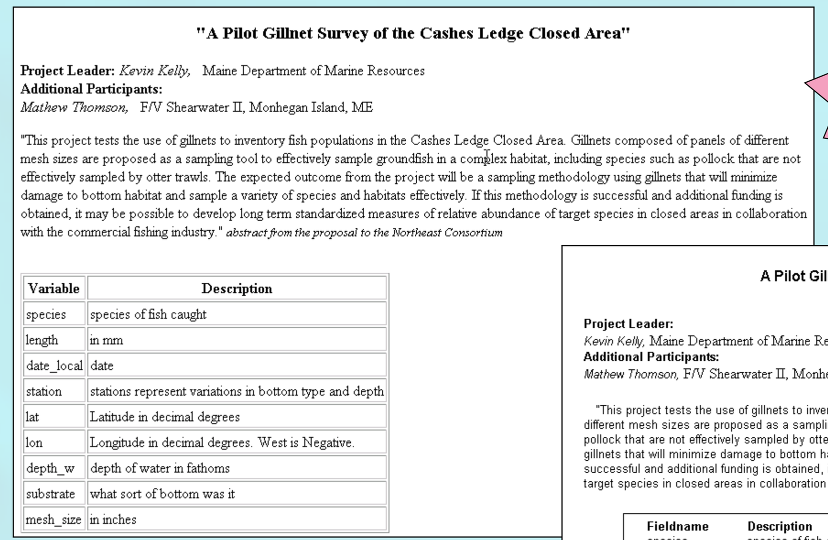
Same plot of Stellwagen Bank Calanus finmarchicus abundance rotated.

Interoperability

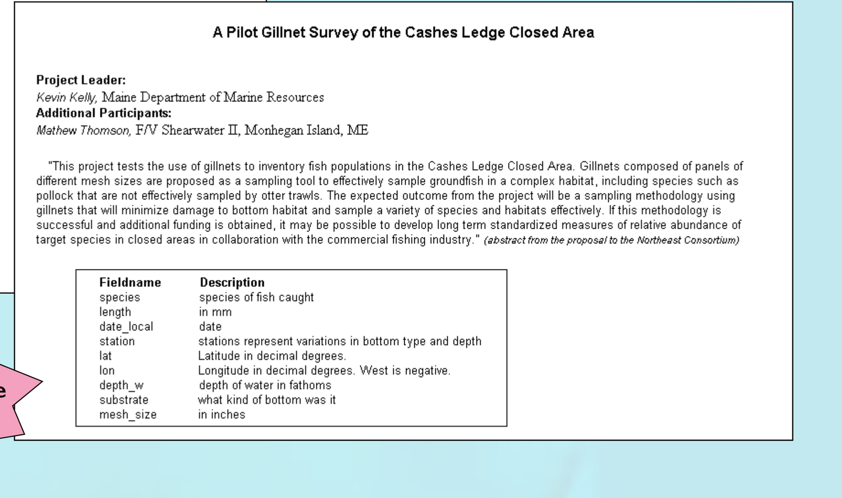
Much national and international effort is directed toward data interoperability. There are a number of different projects designed to make one's data sets compatible and functional with other people's data. It is also clear that data interoperability is not an easy task, nor is there one way, or even a few ways, of accomplishing it. For example,

- SURA SCOOP Project – The Southeastern Universities Research Association's (SURA) Coastal Ocean Observing and Prediction Program is "helping to create an Integrated Ocean Observing System (IOOS) for the Nation by providing some of the IT glue that will unite the coastal science community." They are developing the OpenIOOS Interoperability Test Bed as a "collaborative, multi-institutional initiative to demonstrate and advance data integration and interoperability among coastal ocean observing systems (both academic and government) across the U.S." (Texas A&M, 2006)
- Intergovernmental Oceanographic Commission (IOC) and the International Council for the Exploration of the Sea (ICES) are cooperating in the development of a marine extensible markup language (XML) and have formed the ICES-IOC Study Group on the Development of Marine Data Exchange Systems using XML. See http://www.iode.org/ and http://www.marinexml.net/.
- The Ocean Biogeographic Information System (OBIS, see http://www.iobis.org/) is the information component of the Census of Marine Life (CoML, see http://www.coml.org/coml.htm), a growing network of researchers in more than 45 nations engaged in a 10-year initiative to assess and explain the diversity, distribution, and abundance of life in the oceans - past, present, and future. OBIS is a web-based provider of global geo-referenced information on marine species. Their efforts focus on providing a standardized interface to several database programs, such as Oracle and MySQL, using DiGiR (Distributed Generic Information Retrieval, see http://digir.net/)
- The Monterey Bay Aquarium Research Institute (MBARI) is involved in several projects: 1) Shore Side Data System (SSDS) makes extensive use of XML to describe their metadata. 2) SensorML provides standard models and an XML encoding for describing any process, including the process of measurement by sensors and instructions for deriving higher-level information from observations. Processes described in SensorML are discoverable and executable. All processes define their inputs, outputs, parameters, and method, as well as provide relevant metadata. (MBARI, 2006a). 3) The Marine Metadata Interoperability (MMI) project wants to make it easier to work with marine science data. The project's work is aimed at "scientists and their data management staff who want the best advice and materials for doing data management right. The MMI site identifies best practices to make … science data easy to distribute, advertise, reuse, and combine with other data sets. (MBARI, 2006b)
- The Marine Biological Laboratory is working on XML-based ontology and classification systems.

In addition to providing OGC compliant Web Mapping Services and Web Feature Services from the JGOFS/GLOBEC data management system, our efforts focus on enabling automated access to our metadata. Currently, our metadata is stored in text files that a person is able to read easily via a web page. However, this approach does not provide for access by other computer systems since the data are not organized or "tagged" for automated interpretation. To accomplish this goal, we are in the process of converting all of our human readable metadata files into machine-readable XML coded files. When suitably reformatted, these XML files can look like the original, human readable files, but also provide the necessary structure and encoding for other computer software to directly "read" and interpret this information.



Human readable Metadata



Human and machine readable Metadata

Conclusion

The JGOFS/GLOBEC data management system provides:

- A simple and extensible way to serve one's ASCII and image data via any standard browser..
- New enhancements to improve data discovery and data interoperability include:
  - Incorporating Open Geospatial Consortium standards;
  - Computer readable, XML-based metadata information; and
  - Enhanced options for visualizing the data using MapServer and GeoZui3D.

References

Chandler, C.L. 2006. Live Access Server Description, http://usjgofs.whoi.edu/datasys/LAS-descr.html

Edgington, Theresa, Beomjin Choi, Katherine Henson, T.S. Raghu, and Ajay Winze, Adopting Ontology to Facilitate Knowledge Sharing, Communications of the ACM, November 2004, Vol 47, No. 11.

Flierl, G.R., J. K.B. Bishop, D.M. Glover, S. Paranjpe. 1992. JGOFS Data System Overview, http://globec.whoi.edu/globec-dir/doc/datasys/jgsys.html

Groman, R. C. and P. H. Wiebe. 1997. Management of Biological, Physical, and Chemical Data Within the U.S. GLOBEC Program, in Proceedings of the International Workshop on Oceanographic Biological and Chemical Data Management, May 20 - 23, 1996; NOAA Technical Report NESDIS 8, February 1997.

Monterey Bay Aquarium Research Institute (MBARI). 2006a. http://vast.uah.edu/SensorML

Monterey Bay Aquarium Research Institute (MBARI). 2006b. http://marinemetadata.org

Monterey Bay Area Workshop on Data Management and Visualization, 2003. http://www.mbari.org/iag/workshops/dmv/index.html

Pacific Marine Environmental Laboratory, NOAA. 2006. http://ferret.wrc.noaa.gov/Ferret/LAS/

Texas A&M University. 2006. SURA SCOOP Project. http://www1.sura.org/3000/3310_Scoop.html

Acknowledgments

We would like to thank Mark Benfield, Louisiana State University, for the use of his Stellwagen Bank abundance data images, Richard Taylor for use of the MapServer image, and Colin Ware, University of New Hampshire, for use of the GeoZui3D starting page image. This work is supported by the NSF Ocean Sciences Division, NOAA Coastal Ocean Program and UNH COOA.