Exploring Ocean Data

James G. Bellingham
Monterey Bay Aquarium Research Institute
7700 Sandholdt Road
Moss Landing, CA
1-831-775-1731
jgb@mbari.org

Mike Godin
Monterey Bay Aquarium Research Institute
7700 Sandholdt Road
Moss Landing, CA
1-831-775-2063
godin@mbari.org

ABSTRACT

In fall of 2004, we met Jim Gray and began to converse about the data needs of ocean scientists. The conversations ultimately led to the development of a unique portal for exploring multidisciplinary data sets, which we call the Metadata Oriented Query Assistant (MOQuA). At the time, we were working with an extremely rich data set which included measurements from ships, satellites, aircraft, moorings, and a variety of underwater robots. The data set also included output from both atmospheric and ocean models. We initially made the data available via a state-of-the-art data server. However, serious users did not use the server, instead approaching us for copies of the relevant portions. Our experience convinced us that we needed a far more capable portal, but framing the seemingly divergent needs of ocean scientists in a way which could be satisfied via an intuitive interface was challenging. In the course of many conversations and interactions with Jim, we realized that we needed to structure the interaction around questions rather than visualizations, and this simple insight lead to the development of MOQuA system.

Categories and Subject Descriptors

H.2.8 [**Database Applications**] – data mining, scientific databases

General Terms

Management, Experimentation, Human Factors, Standardization

Keywords

Oceanography, data exploration, heterogeneous observations

1. INTRODUCTION

The ocean is a complex, highly interconnected environment. Scales of spatial variability range from microscopic to thousands of kilometers. Temporal variability is linked to spatial scales, and spans from fractions of a second to thousands of years. Adding to this complexity, a full description of the ocean must include not just physical characteristics, such as temperature and currents, but chemical, biological, and even geological parameters. A variety of linkages, some known, and probably many more unknown, transmit variability of one parameter, in one location, to another parameter in a quite different location. For example fisheries off of Peru are strongly regulated by interactions of wind and ocean in the central Pacific (the El Nino Southern Oscillation phenomena).

Unfortunately, the ocean is extremely sparsely sampled, compared to its scales of variability. While satellites can image the sea surface, they are not capable of directly measuring interior properties of the ocean. To measure interior properties a range of methods are used, all revolving around either measuring properties with *in situ* sensors, or obtaining seawater samples and returning them to a laboratory. The advent of robotic systems has dramatically reduced the cost of making simpler measurements in the ocean, and thus allowed an increase in observations[1]. However, even the most intensive field programs fall short of synoptic characterization of the underwater environment.

The twin realities of measurements which are sparse compared to the phenomena under study, yet which are numerous in both number and type, frame the challenge for ocean data systems. An ocean scientist might work with atmospheric observations, surface currents measured by radar, time series observations from a mooring, and measurements of chlorophyll fluorescence made by an underwater robot in the process of understanding the source of a harmful algae bloom. The data system must not just serve data from each of the disparate sources, but must help the oceanographer find the small portion of the archive with relevant information.

This paper describes a metadata oriented data exploration tool we developed to assist ocean scientists explore data sets containing information from a wide range of sources. Jim Gray's insight and guidance was instrumental in shaping our efforts. In the sections below, we provide brief overviews of the types of data involved, our initial approach to providing a portal, the transformational conversation with Jim, and the resulting portal.

2. OCEAN OBSERVATIONS

2.1 Observation Platforms

Observations of the ocean are made by a wide variety of systems, each with their own particular data characteristics. It is not unusual for ocean scientists to rely on data from more than one source to gain a full understanding of a particular process.

Examples of data sources include:

- Ships may collect data along the surface, at profile stations, or along a saw-tooth pattern if a tow-fish is dragged behind the ship. Typical ship deployments last between a day and a

Jim Gray Tribute, May 31, 2008, Berkeley, CA, USA.

week. However, it is not uncommon for ships to regularly repeat their deployments, creating multi-decade time-series.

- Moorings are nominally fixed in location, making the same measurements for long time periods. Some ocean mooring have been making regular measurements for decades.
- Current sensing radars regularly measure the surface currents for a fixed region of the ocean. Like moorings, these systems are capable of making very long-duration time series measurements.
- Surface floats drift with currents, making measurements along the ocean surface.
- Profiling drifters spend most of their time drifting with currents at a set depth. Periodically these systems will rise to the sea surface, where they report measurements by satellite, and then return to depth. While drifters can operate for years, their path, and therefore the areas that they measure cannot be controlled.
- Gliders change their buoyancy to rise and sink through the water column, using wings to translate vertical motion into horizontal motion. These vehicle will periodically pause at the sea surface to report measurements and receive new commands. Gliders typically collect data for months.
- Propeller-driven autonomous underwater vehicles (AUVs) collect data along any 4-dimensional path through the water column, typically for periods of about a day. AUVs typically travel about five times faster than a glider, and carry more complex and power consumptive payloads.
- Aircraft collect data along 4-dimensional paths through the atmosphere, as well as surface measurements of the ocean. Typically, aircraft collect a very dense data set, as their cost of operation limits them to occasional flights.
- Satellites make measurements of large areas of the ocean, but typically not very often (for example once or twice per day), and typically along "swaths" that do not repeat.

2.2 Observation Systems

Ten years ago, a typical oceanographic field program was structured around measurements made from a single ship, which were only fully analyzed in the months and years after the cruise was complete. The advent of robotic systems has enabled experiments using distributed arrays of mobile assets (figure 1) which can provide three dimensional views of the ocean interior over limited regions (figure 2). The mobility of these assets creates both opportunities and challenges. The opportunity stems from the ability to adapt observations to changes in the environment. The challenge stems from the need to manage a large number of assets, leading in turn to a greater emphasis on analyzing observations immediately to support decision making. Thus the evolution in observational capability has lead to the need for a communications and cyberinfrastructure which allows real-time collection and analysis of observations.

3. DESIGNING THE PORTAL

3.1 The Starting Point

The traditional model for oceanographic data exploration has been a two-step process that begins with a scientist downloading an entire data set from a science archive to their local computer and ends with the data set being searched by a piece of analysis software on the scientist's computer. This is the only method of exploring many ocean data collections today. As datasets grow to terabyte and petabyte sizes, such a model becomes unwieldy [2].

Moving beyond the download and filter model for accessing oceanographic data, two products stand out: the Live Access Server (LAS) [3] and Dapper Data Viewer (DChart) [4]. They are both open-source web applications made available through the National Oceanic and Atmospheric Administration (NOAA) Pacific Marine Environmental Laboratory. Each allows one to select a data set and variables of interest, and both use form-based and graphical tools to allow users to interactively select a region of interest to query for (and return) data only within that region.

LAS is designed to serve data that is either stored in Network Common Data Form (NetCDF) files or data that can be accessed via the Open-source Project for a Network Data Access Protocol (OPeNDAP) protocol. LAS is highly flexible, and if a relatively static dataset is being served, one can configure LAS to display where data was collected while the selection region is manipulated. It also allows for data to be cataloged in configurable hierarchies.

DChart allows one to visualize and download in-situ oceanographic or atmospheric data from a Dapper OPeNDAP server. Features include an interactive map that is draggable, an in-situ station layer that allows users to select data stations, and a plot window that allows one to plot data from one or more stations. Three plot types are supported (profile, property-property, and time series) and users can interact directly with the plot to pan or zoom in and out.

A weakness in both of these systems is the fundamental assumption that one knows which data sets they are interested in when they start their data search. For example, asking a simple question like "What systems were measuring temperature in Monterey Bay on August 10th, 2004?" requires an investigator to access every data repository that contains Monterey Bay temperature measurements independently, and could take hours.

3.2 Asking Questions

Once the data is in hand, most scientists begin an exploration of a data set by producing plots. Consequently, a conversation with a scientist about a data portal naturally gravitates to a discussion of the types of graphics which will be generated. On the part of the engineer, there is the desire to isolate the key features required that will satisfy the majority of users. On the part of the scientist, there is an endless variety of ways in which the data can be presented. In engineering terms, far from converging on a common set of needs, the framing the discussion around visualization leads to a never ending stream of requirements.

Perhaps the most important single lesson we learned from Jim was that a far more productive discussion is to identify the key types of questions scientists will want to address with the data[5]. Through spring of 2005 we had a series of meetings and email interactions with Jim in which he explained the thought processes behind the creation of the highly successful Skyserver portal to the Sloan Digital Sky Survey. Together with Jim, we worked through a similar chain of logic for an ocean data portal. Jim probed the way in which oceanographic data is explored in a series of email interactions. The ongoing analysis of the 2003 data provided a variety of concrete examples we could share with Jim. Ultimately Jim sent a mock-up of a portal (figure 3) along with the text description below:

"1. I think the world wind viewer Keith is cooking up (or some derivative of it) will be the way to get oriented in space. One can ask for "tracks" or "platforms" or "footprints" (for satellite or survey data) be rendered as layers above the backdrop. These layers could be selected from a list and turned on/off (world wind has a prototype for that). Lets call that the LOCATION window. (upper left windows in screen shot below). Other windows can send the location window events and as you move around the location window it can send events to other windows.

This "BRUSH" effect is a fairly intuitive way to explore multidimensional data.

As you scroll through the "DEPLOYMENT SCHEDULE" window, it would affect what is rendered in the LOCATION window. (I show a time limit on the deployment window afffcting the other windows and a track on the LOCATION window affecting the others)

2. Other windows can have plots of x vs y and x vs y vs z for any x,y,z you care to define.

When those dimensions are spatial or temporal there is an obvious backdrop but generally they are not.

The pane at right controls which layers are visible in each window.

3. So now we are into defining x,y,z. They can be a database query but more likely they are the output of some analysis tool. DB queries are "easy" but require the scientist to think at a low level and speak a funny language. We will start with that and that will be there as an escape hatch in case the analysis tool does not do what is needed, but... The goal is for the DB to be hidden."[6]

The reference to the "world wind viewer Keith is cooking up" refers to the work of Keith Grochow, who has been developing a three-dimensional georeferenced ocean workbench [7]

3.3 Linking Questions to Metadata

The template for a ocean data portal which Jim described gave us new directions to pursue, and helped solve some other problems with which we had been grappling. Although the portal had been a primary focus of effort, we had also been working to resolve considerable confusion regarding naming conventions for environmental variables. The AOSN 2003 data set was generated by more than a dozen research groups representing different scientific and engineering disciplines. Not surprisingly, the different groups conformed to different ontologies, or in some cases, made up their own. A query oriented portal provided us an ideal framework for a more intuitive approach to interacting with named variables and assets.

As a first step toward developing a multidisciplinary data portal, we focused on creating a tool that combined exploration of the variable space with the more traditional spatial and temporal exploration. We call this tool the Metadata Oriented Query Assistant (MOQuA)[8]. The internet browser-based MOQuA user interface provides visual selection of data sets, variables of interest, and regions of interest from multiple data collections simultaneously (see figure 4). For example, if one is exploring a geospatial-temporal data sets and selects a 4D region of time and space, MOQuA highlights the data sets and variables that have data in that realm. If one selects one or more variables, MOQuA highlights the data sets that contain those, and a representation of the data appears on a selection map, on a time line, and on an altitude selector. Likewise, selecting data sets highlights the variables contained in the data sets.

4. LESSONS AND FUTURE DIRECTIONS

The mockup described by Jim Gray in his May 2005 email has only been realized in small part. However, our work to create portal elements have resulted in important insights regarding the organization of metadata to support data exploration. For example, hierarchical descriptions allow a more natural interaction with query tool. The most important lesson is the value of registering data against multiple ontologies. Not only does this allow users familiar with a particular ontology to explore data in a familiar framework, but different ontologies are structured to reflect different ideas about what is important in the data, and thus support different types of exploration.

In the coming year, we plan on adding data preview windows to MOQuA, which will move it closer to the vision outlined by Jim. We also are working to make MOQuA metadata pivot lists selectable. In effect, this should let users to query on other properties, such as elements of the data provenance, and user-assigned tags and ratings. Ultimately we see this tool as being useful for not just ocean data, but for any scientific data enterprise involving interdisciplinary data exploration.

5. ACKNOWLEDGMENTS

We owe thanks to Ed Lazowska for many deeds, but foremost for introducing us to Jim Gray and encouraging many valuable conversations. Development of MOQuA was supported by the Office of Naval Research (ONR) under grant N000140210856 and by the David and Lucile Packard Foundation.

6. REFERENCES

[1] Bellingham, J. G., and Rajan, K., "Robotics in Remote and Hostile Environments," *Science*, Vol. 318. no. 5853, pp. 1098 – 1102, November 2007. doi: 10.1126/science.1146230

[2] Gray, J., Liu, D. T., Nieto-Santisteban, M. A., Szalay, A. S., Heber, G., DeWitt, D., "Scientific Data Management in the Coming Decade," ACM SIGMOD Record, v.34 n.4, p.34-41, December 2005 [doi: 10.1145/1107499.1107503]

[3] Foley, D.G , and de Witt, L.M. , "The OceanWatch Live Access Server," *Proc. MTS/IEEE Oceans 2005*, Washington, DC: 2005. doi: 10.1109/OCEANS.2005.1640012

[4] Sirott, J., "DChart: A Remote Scripting Web Application for In-Situ OPeNDAP Data," *22nd Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Atlanta, GA, February 2006.

[5] Gray, J. and Szalay, A., "Where the Rubber Meets the Sky: Bridging the Gap Between Databases and Science," *IEEE Data Engineering Bulletin*, Vol. 27, No 4, pp.3-11, October 2004.

[6] Personal communication: Jim Gray, April 28, 2005.

[7] http://www.cs.washington.edu/homes/keithg/oceans.html

[8] Godin, M. A. and Bellingham, J. G., "Data Exploration for Multidisciplinary Research," *Oceans 2007*, pp. 1-4, Sept. 29 2007-Oct 4 2007. [doi: 10.1109/OCEANS.2007.4449264]

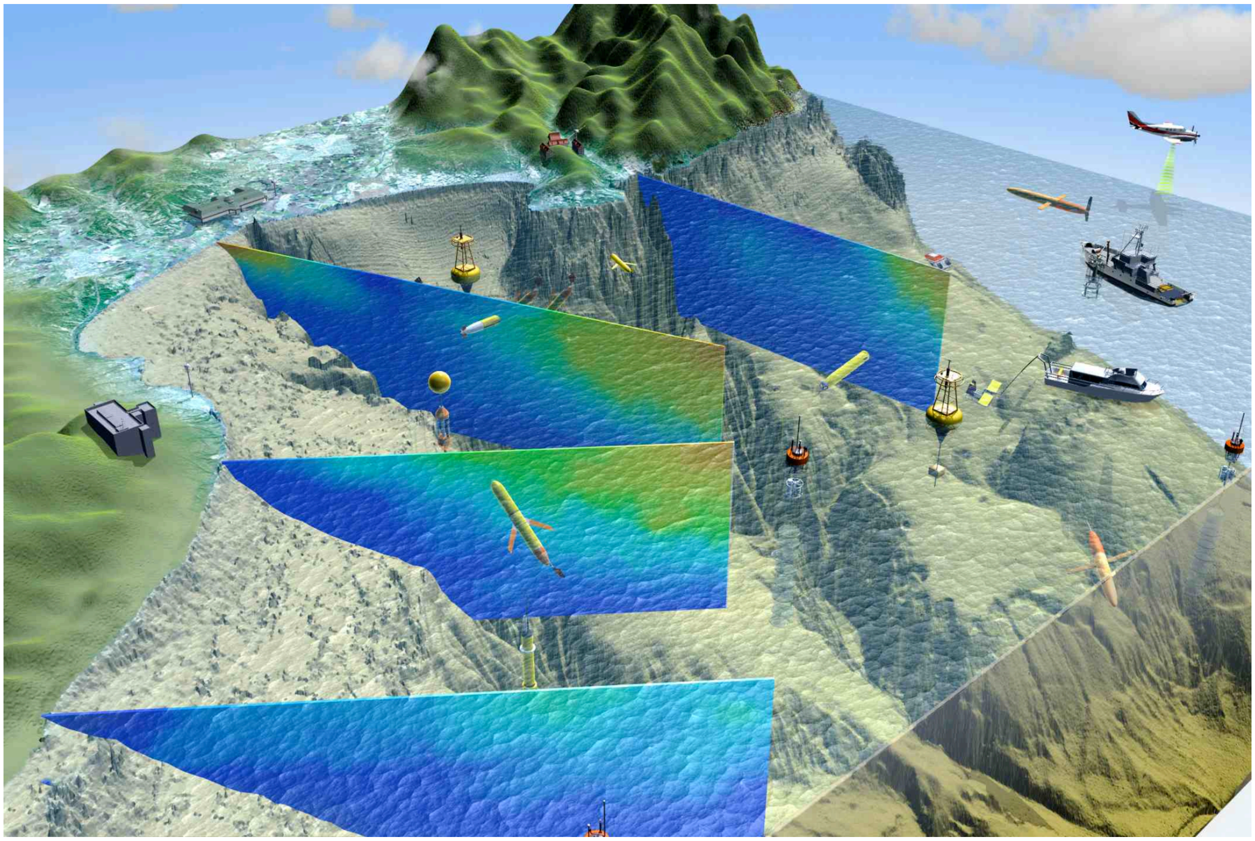


Figure 1: This figure illustrates an intensive field program, increasingly characteristic of ocean science. The region depicted is Monterey Bay during the Autonomous Ocean Sampling Network 2003 experiment. A wide range of observational tools are illustrated, including ships, aircraft, moorings, drifters, and a variety of underwater vehicles. In the 2003 field program, these assets were tied to a data system on shore, which in turn made the data available for assimilative models, which predicted future conditions in the Bay. The colored vertical sections extending from shore show water temperatures as measured by underwater vehicles, showing colder (blue) water being brought to the surface along the coast, displacing warmer (red) surface water offshore. This is characteristic of wind-driven upwelling along much of the California coast.

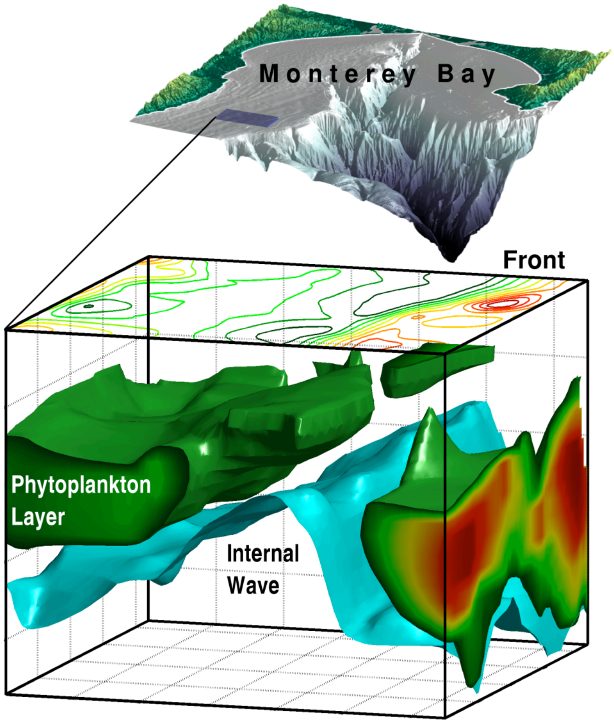
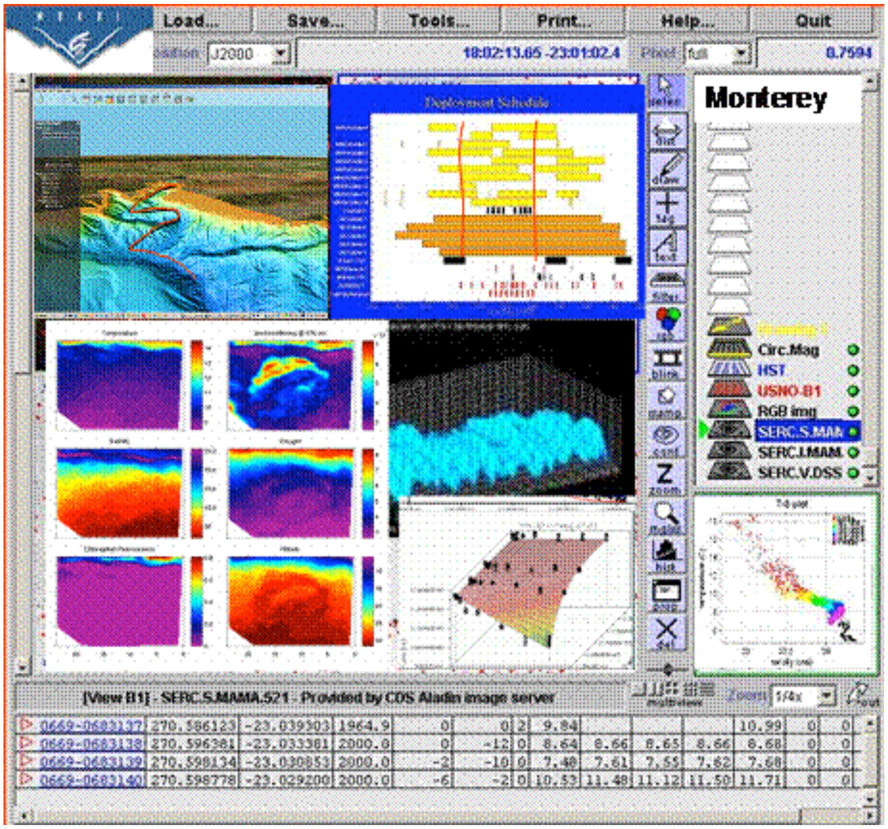


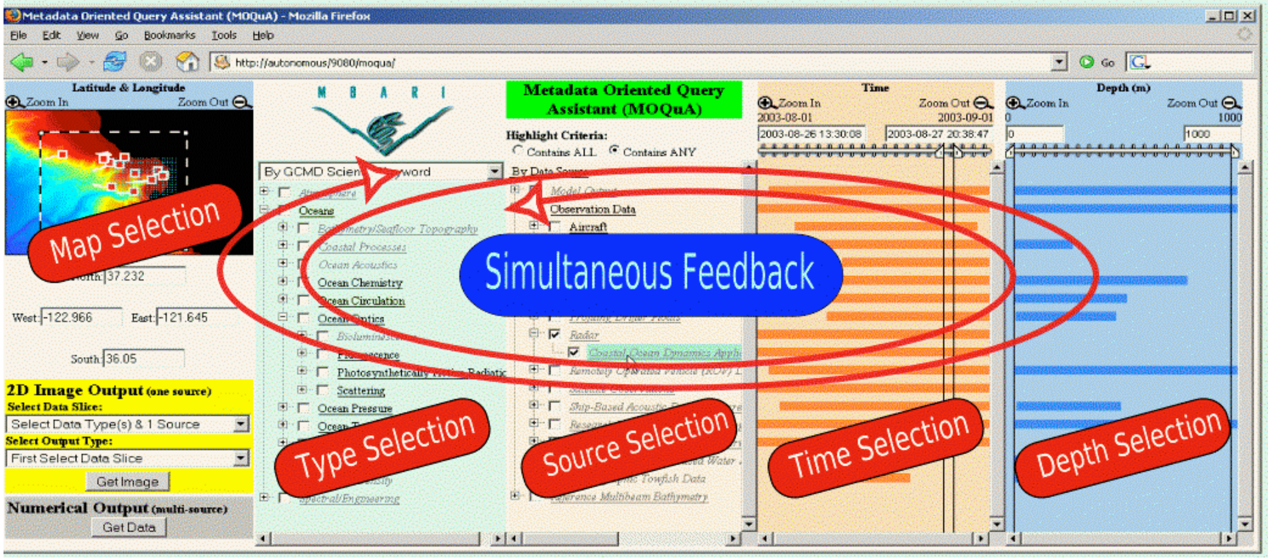
Figure 2: A three-dimensional image of interaction of physical and biological processes, mapped by an Autonomous Underwater Vehicle. The green volume depicts a phytoplankton layer while the underlying cyan surface shows deflection of a surface of constant density by an internal wave. The volume shown is 6.5 by 2.5 km in horizontal extent and 23 m in depth.



[View B1] - SERC.S.MAMA.521 - Provided by CDS Aladin image server

0669-0683137	270.586123	-23.039303	1964.9	0	0	2	9.84				10.99	0	0
0669-0683138	270.596381	-23.033381	2000.0	0	-12	0	8.64	8.66	8.65	8.66	8.68	0	0
0669-0683139	270.598134	-23.030853	2000.0	-2	-10	0	7.48	7.61	7.55	7.62	7.68	0	0
0669-0683140	270.598778	-23.029200	2000.0	-6	-2	0	10.53	11.48	11.12	11.50	11.71	0	0

Figure 3: This is the mockup of a portal created by Jim Gray to illustrate what a data workbench might look like for oceanographic data. Jim's idea of how this would be used is given verbatim in the text, but in brief involves using a 3-d view of trajectories over bathymetry (upper left), selecting time from a graphic of deployments (middle right), and allowing plots of variables against each other. The objective is to let the user frame a database query by interacting with the graphics, without their having to "speak a funny language."



http://autonomous/9080/moqua/

Highlight Criteria: ○ Contains ALL ● Contains ANY

North: 37.232
West: -122.966 East: -121.645
South: 36.05

Select Data Slice: Select Data Type(s) & 1 Source
Select Output Type: First Select Data Slice

Time: 2003-08-01 – 2003-09-01; 2003-08-26 13:30:08 – 2003-08-27 20:38:47

Depth (m): 0 – 1000; 0 – 1000

Figure 4: MOQuA interface, showing multiple selection screens for exploring data. Five different interacting selection panes are shown in this view. From left, there is a map selection pane, which allows the user to specify the region of interest, and to view the spatial extent of selected sources. Next, there is a measurement selection pane, which shows types of observations and model results available, both for the entire data archive, and for the selected data sources. The middle selection pane allows one to select observation and modeling sources. At the right are panes which allow selection of time and depth periods of interest. There is a high degree of interactivity between the various panes, for example the sources listed in the source pane are underlined when they are present in the map, time, and depth selection panes. Similarly, the sources are shown in bold when they provide measurements selected in the type pane.