Geospatial Access and Data Display Adds Value to Data Management at the Biological and Chemical Oceanographic Data Management Office

M. Dickson Allison[1] and Charlton Galvarino[2]
[1]Woods Hole Oceanographic Institution, Woods Hole, MA
[2]Second Creek Consulting, LLC, Columbia, SC

In 2006, the Biological and Chemical Oceanography Data Management Office (BCO-DMO) was funded to manage all data supported by the Biological and Chemical Oceanography Sections and (later) the Division of Polar Programs Organisms and Ecosystems Section of the U.S. National Science Foundation.

BCO-DMO serves a wide variety of oceanographic related data, including but not limited to biological, chemical and physical oceanography measurements. We routinely deal with CTD data, biological abundance, meteorological data, nutrients, pH, carbonate chemistry, PAR, sea surface temperature, heat and momentum flux, sediment composition, trace metals, primary production, and pigment concentration measurements, and with images and movies. In addition to a text-based interface, we offer an OGC-compliant geospatial interface (MapServer[1]) to enable the user to find the data, evaluate its fitness-for-purpose, and download and re-use the data for his or her own analyses.

This task speaks to new data but since the data management office of BCO-DMO was formed in 2006 from the merging of the U.S. JGOFS (1984-2003) and the U.S. GLOBEC (1994-2006) Data Management Offices, there were thousands of legacy datasets to be incorporated into what became the BCO-DMO data management system. Because BCO-DMO is equally devoted to bringing 'dark data' into the light, legacy datasets are still being welcomed.

BCO-DMO works directly with Principal Investigators (PIs) to make their data available on the web. There really are no data format restrictions. (This has serious consequences for geospatial access and will be discussed further below.) There are some best practices[2] PIs are encouraged to follow, and which provide their own reward in terms of speed of data availability.

BCO-DMO manages data that are heterogeneous in the extreme, and therefore one method of geospatial display does not work for all. Because the data are not always the same kind, format, scale, or magnitude, each 'quick look' at the data using the map must be individually tailored. Figures 1 – 5 illustrate this point.

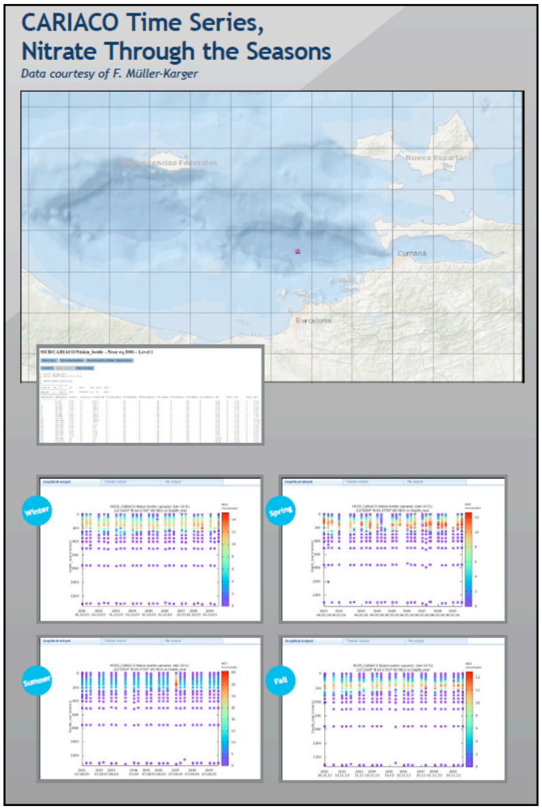

Figure 1: CARIACO project chemistry time-series data at the site indicated on the map. Individual plots illustrate changes in Nitrate concentration by season.

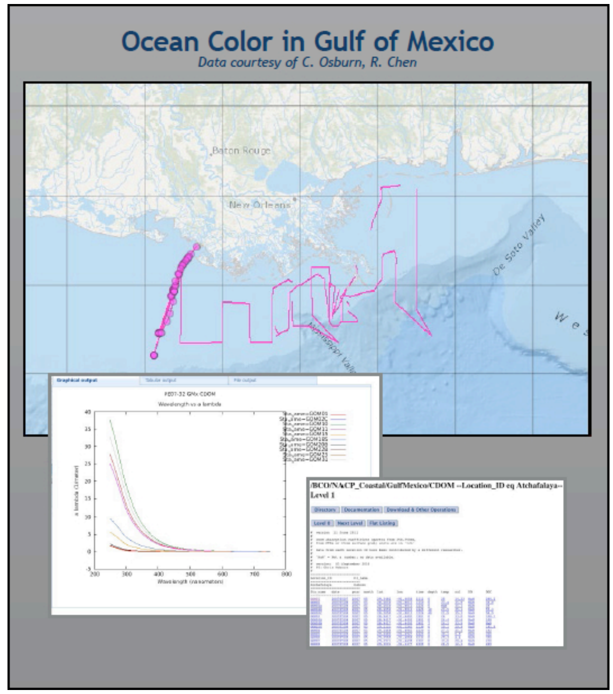

Figure 2. Dissolved Organic Matter changes with latitude in the Gulf of Mexico

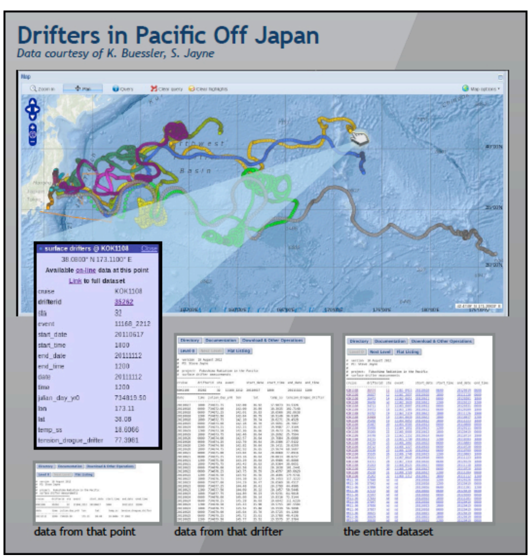

Figure 3. Drifter tracks illustrating particle movement following the Fukushima Earthquake at three different data granularities.

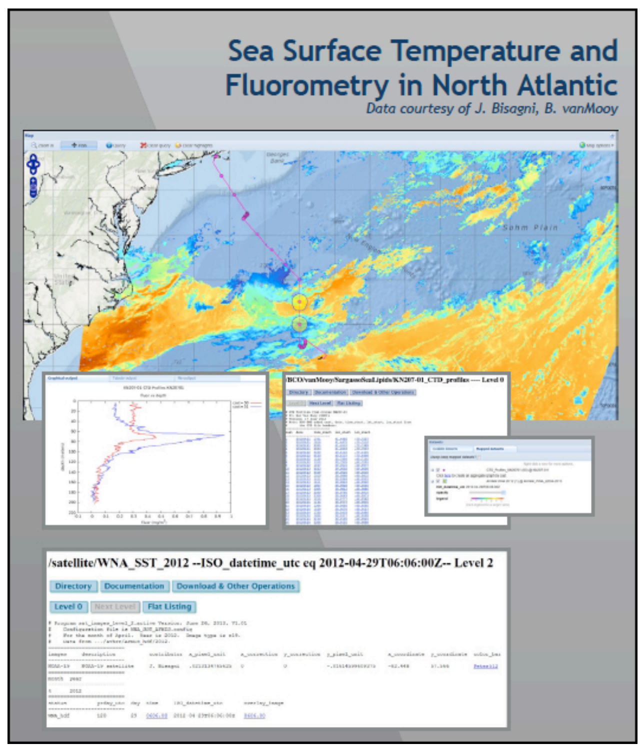

Figure 4. Fluorometry and its relationship to Sea Surface Temperature from two different data sources.

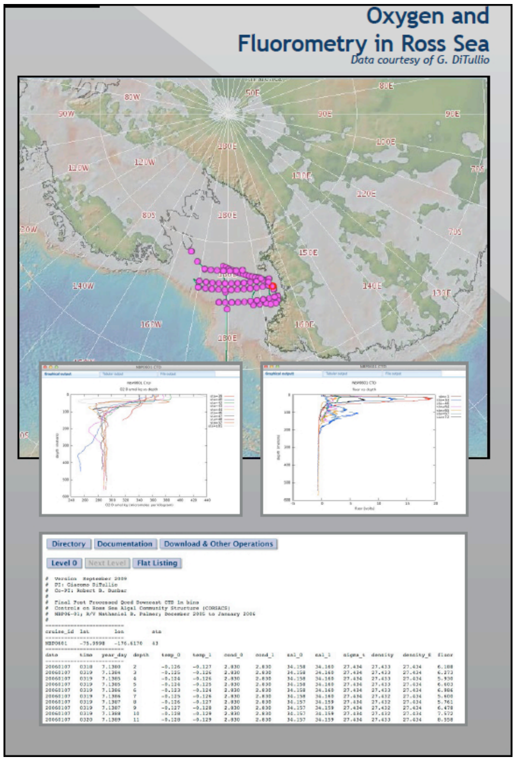

Figure 5. Polar projection to illustrate algal community structure in the Ross Sea

The challenges of providing different displays for different types of data are many. Each dataset must be described for the mapper. This is done via specific mapping parameters assigned by the data managers and added to the metadata associated with each dataset and stored in a MySQL database.

See Figure 6 for a representation of the [Metadata Database Schema](). (Since we converted from Cold Fusion to Drupal in November 2013, the MySQL database now contains a collection of Drupal tables as the source of the RDF triples. The RDF triples are used in our Beta version of an [Advanced Semantically-enabled Faceted Search](), a part of which is illustrated in Figure 7.) Essential mapping information like deployment latitude and longitude are in both the metadata and the data themselves (accessed directly) and indeed could be considered both data and metadata (perhaps along with water depth and time). Once we have collected information about the data and entered the metadata into the system we add additional information about the mapping parameters needed for the mapping and display software.

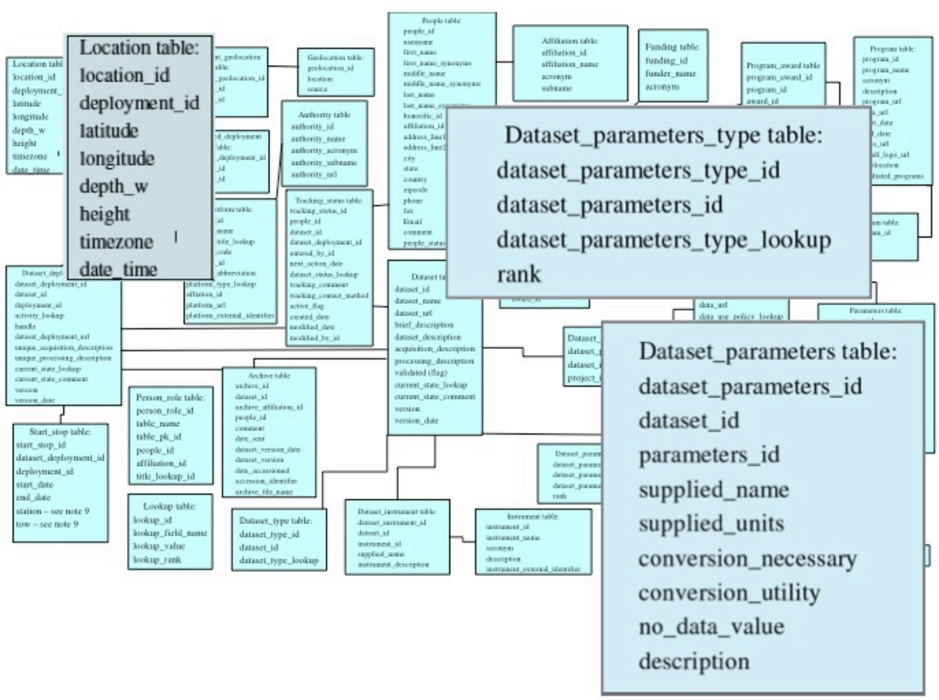

Figure 6. MySQL Database Schema emphasizing mapping instruction tables.

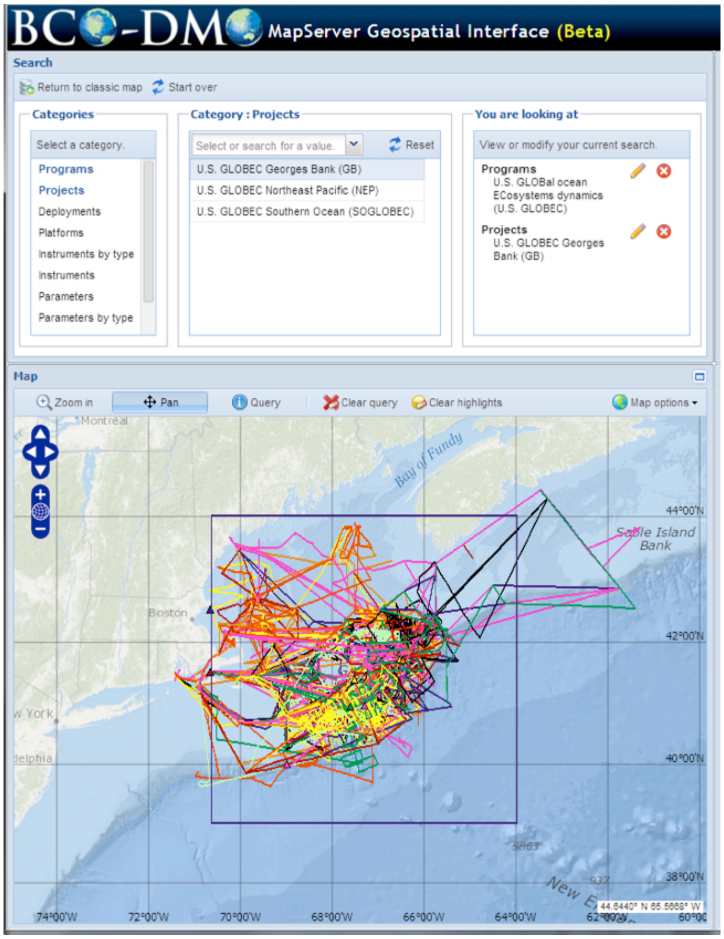

Figure 7. An example of a portion of the Advanced Search (Beta) display showing a semantically-driven search of programs and projects.

The requirements for mapping can then be broken up into two steps: (1) showing the location of the dataset and then (2) displaying the dataset itself using an appropriate graph or plotting format.

1. When a user visits the Mapper, the initial visualization is of many deployments. A deployment is one of: a ship's position during a scientific cruise or 'cruise track'; a fixed sampling location as from a moored buoy; a position of the laboratory where the results were created; or a satellite coverage bounding box. In order to provide this initial visualization, each deployment must report geographic and temporal information. In the case of a cruise track, a series of [longitude, latitude, and time] triples are reported along with a flag indicating this to be a LINESTRING dataset. This metadata database-reported position information is grouped together into one row in a Shapefile and is accessed by the MapServer software to provide the deployment visualization.

2. Now that the deployments have been visualized, a user will likely interrogate them to explore the datasets collected en route (in the case of a cruise track). From the Mapper's perspective, it is only valuable to show one point at any given location. Having overlapping points for a single dataset is meaningless to a user.

a. The Mapper has a finite set of ways it may visualize the data and this is driven by the metadata. Is this dataset a simple point in time and space? Is this dataset a collection of similar data collected at the same point over a period of time? Is this dataset a collection of datasets that vary over time and depth and space? Is granularity important? The answers to these questions determine such things as whether or not a table of data is displayed, a graph is displayed, or a KML file is generated. In the case of KML output, the user is invited to explore the KML data in Google Earth to plunge under the sea to examine specific data points across a wide geographic footprint.

b. The point at which data are displayed as a table or presented as a graph marks a departure from the Drupal metadata database. It is at this point that underlying data sources are accessed directly. Since there are no data submission format restrictions (as mentioned earlier in this document), reading the data directly is challenging. For example, determining the time at which a particular row of data was collected varies from dataset to dataset: in one dataset the time and date may be broken into 6 columns while in another, the time is a UTC string in a single column.

The next challenge is how do we communicate with the user how to make the 'quick look' displays that we hope will add value to the data?

We address the challenge of navigating the mapping options by providing aids in the form of descriptions, tips, and help boxes in as many places as possible, including: Map Option pull-downs, directions in the panel banners, help modules, and consistent use of 'right click' options, among others. However, it has become obvious that the more creativity and software flexibility that is provided to deal with a growing list of data types, the more complex the user experience becomes.

We are always asking: Are we giving the user what they really want? Are we giving them **all** the data about radioactivity levels reaching the West Coast of the US or Deep Water Horizon oil spill hydrocarbons affecting the benthos in the Gulf of Mexico or only what BCO-DMO can make available via area selection ("rubber banding"), searching and downloading? What about Monterey Bay Aquarium Research Institute (MBARI) data or University of Washington data? Can we link to those data and use our own display methodology? And even if we could, would it be at the same scale or precision or quality? There are many reasons for linking geospatial data from multiple sources but there are cautions as well.

References:

[1]http://www.mapserver.org. Originally developed at the University of Minnesota.

[2]BCO-DMO Data Management Best Practices Guide: http://www.bco-dmo.org/data-management-best-practices-guide-0