SEA TECHNOLOGY, vol. 46 (2): 93. (February 2005)

SOAPBOX (invited submission)

Data Management a Top Priority?

(slightly longer than final printed version)

Dr. Dawn Wright has for many years been interested in solving problems related to the management and spatial analysis of oceanographic data, with an emphasis on application issues for GIS. She has completed oceanographic fieldwork in some of the most geologically-active regions of planet, and has had three dives in the Alvin *submersible. Dr. Wright is professor of geography and oceanography at Oregon State University, and holds degrees from UC-Santa Barbara (Ph.D. in marine geology and geography), Texas A&M (M.S. in oceanography), and Wheaton College in Illinois (B.S. cum laude in geology). She serves on the editorial boards of the* International Journal of Geographical Information Science*,* Transactions in GIS*, and* Geospatial Solutions*, and was a member of the National Academy of Sciences' National Needs for Coastal Mapping and Charting Committee.*

This article is about an encouraging trend that I'd like to see continue, where the management of data from instruments and platforms becomes nearly as important as the scientific questions and hypotheses that drove the collection of the data in the first place. I hope that it is indeed a trend because it has great implications for how much "bang for the buck" that we really derive expeditions and *in-situ* deployments and observatories. About a decade ago several "voices in the wilderness," including the early adopters of marine geographic information systems (GISs), argued that the expense of going to sea alone justifies the development or implementation of systems to manage the resulting data. Often in the past with oceanographic field programs, regardless of size, the scientific questions were singularly foremost, but the accompanying data management

issues were ignored until the end. It was then revealed that that those problems were so daunting, that it was very difficult to actually assimilate, analyze, and distribute the data in order to answer or revisit those great scientific questions. Happily we have experienced a slow transition from Scenario 1 (going to sea and collecting huge quantities of data that remained proprietary or undocumented, and hence unusable after initial collection); to Scenario 2, where there was more of a willingness to share and document data after the fact (partially through an improved understanding of the importance of metadata); to Scenario 3, where data management and distribution issues are now being considered at the initial planning stages of the largest of coordinated projects such as the Integrated Ocean Observing System (IOOS) or the Integrated Ocean Drilling Program (IODP).

Not only is this critical for the effective use of the data by more people in order to solve the basic and applied science questions (or to make ocean policy decisions), but also because the management is becoming recognized as a "science" itself. This "science" is often in the realm of information technology or geographic information science, two interdisciplinary fields that some oceanographers and marine technologists are now shifting into. Such shifts should indeed continue, as evidenced by the recent NSF/ONR sponsored report, *An Information Technology Infrastructure Plan to Advance Ocean Sciences* by the Ocean Information Technology Steering Committee (http://www.geo-prose.com/oiti/). There are many robust, cutting-edge questions in the realm of data management, or more specifically informatics, that are waiting to be solved, and focusing on ocean instrumentation, platforms, and oceanographic science and policy makes them even more exciting and challenging. These cutting-edge topics include spatial ontologies (the formalization of concepts and terms used in fieldwork, research, and industry, and thus a data "language," often in the form of catalogs, glossaries, thesauri, etc.), semantic interoperability (distinguishing between data languages and mapping them to common language so that data sets can be found and used interchangeably), data mining and knowledge discovery (finding patterns and subtle relationships in data sets, and deriving a resulting interpretation), data fusion (combining and integrating data sets), and data modeling (conceptual formalization of how data are collected, stored, and organized for effective use by a computer application).

There are many large oceanographic data management efforts in progress to watch and draw from. For example, the ArcGIS Marine Data Model project (http://dusk.geo.orst.edu/djl/arcgis; http://support.esri.com/datamodels) is in the realm of data models, spatial ontologies, and data fusion. Next year it will release a generic template and set of user case studies for taking fuller advantage of the most advanced manipulation and analysis capabilities of ArcGIS, particularly its support of more complex rules that can be built into its geodatabases, and of objects with not only attributes, but behavior. A further goal is the support of existing data standards to help simplify the integration of data at various jurisdictional levels for large "enterprise level" GIS projects, including those in

support of ocean observatories. For users, the model provides a basic template for implementing GIS projects (e.g., inputting, formatting, analyzing, and sharing data); for programmers, it provides a basic framework for writing program code and maintaining applications that will be shared throughout the enterprise.

While the ArcGIS Marine Data Model project seeks to promote the interoperability of data and software for user in ocean science and resource management, the new Marine Metadata Interoperability project (http://marinemetadata.org/) is true to its name, with a focus on various effective ways for documenting that oceanographic data. In the coming months this collaborative will provide the marine data management community with general information, standards, ontologies, tools, “cookbooks,” working examples. New registrations to the site are now being accepted.

There are many other projects and initiatives too numerous to mention all of them here. Notables include the long-standing Distributed Oceanographic Data System (DODS), the Ocean Biogeographic Information System (OBIS), NEPTUNE, the new RIDGE/MARGINS Data Management systems, au- and dbSEABED in Australia and the U.S., CELTNET in Ireland, etc. Hats off to the large observatory efforts where data management is finally being talked about and strategized up front, sometimes even before instruments and platforms are in place, and hence before the incoming data streams become the proverbial fire hose to be drunk from, with the accompanying quagmires to be fixed after the fact. Initial data management considerations are still important for smaller projects as well (akin to balancing one’s checkbook every month before things get out of control). Indeed, might this also affect the kinds of questions that one might ask (i.e., as much as the tools and technologies are driven by wanting to understand how the oceans work, so new process questions may be posed because of the informatics)? In the end I believe it comes down to a fundamental change in culture, not only in terms of how we document, share, and collaborate with a data set, but also in how we regard the importance of scientific questions inherent in the management of the data set itself.