Building a Global Plankton Database: Eight years after Hamburg 1996

Todd D. O'Brien

Marine Ecosystems Division – National Marine Fisheries Services - F/ST7
1315 East-West Highway – SSMC III, Silver Spring, Maryland 21044 - USA
E-mail: Todd.OBrien@noaa.gov

Abstract

The "International Workshop on Oceanographic Biological and Chemical Data Management" held in Hamburg (Germany), 1996, produced a listing of suggested metadata for plankton data management. In the eight years that followed this meeting, the efforts and experiences of adding plankton tows to the World Ocean Database profile database made it clear that there was more to building a useable plankton database than just putting plankton data and metadata into a database.

Keywords: NMFS-COPEPOD; Plankton Database; Zooplankton data; Phytoplankton data; Abundance data; Biomass data; Composition data; Quality control.

Introduction

At the "International Workshop on Oceanographic Biological and Chemical Data Management" (Hamburg - Germany, 1996), Linda Stathoplos and Todd O'Brien presented their initial efforts to include plankton tows in the World Ocean Database's profile-based architecture. The Workshop produced a listing of suggested "metadata", ancillary information about the data collection and processing methods, which should be co-stored with the plankton data to ensure usability. For eight years this metadata listing was used as a general guideline for what ancillary sampling information to store in the World Ocean Database, but it became obvious that there was more to building a useable plankton database than just putting plankton data and metadata into a database.

World Ocean Database 1998

Plankton data continued to be added to the database for two years after the 1996 workshop. Linda Stathoplos left for private industry, and the author took over leadership of the effort. In 1998, this global collection of plankton data first became public with the release of World Ocean Database 1998 (WOD98, Fig. 1, Conkright *et al.,* 1998).

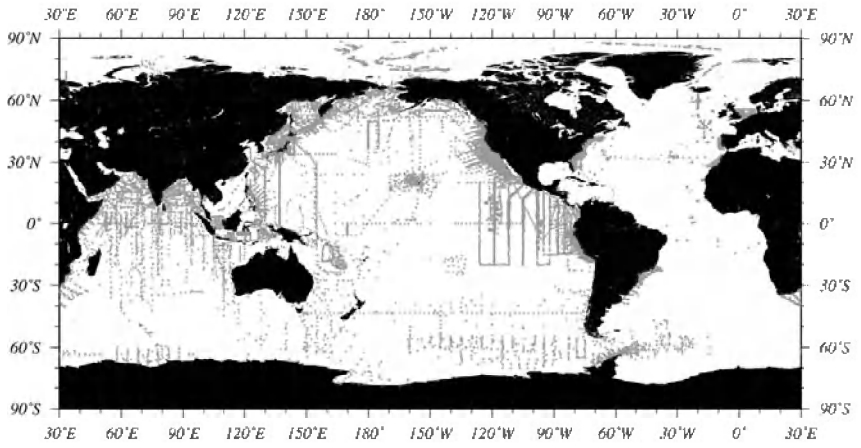

Fig. 1. Plankton tows present in World Ocean Database 1998.

While the WOD98 plankton data followed the Hamburg 1996 metadata guidelines, they were nearly impossible to use because of the complexity of the biological data itself, coupled with the complexity of the World Ocean Database profile-based data format. The plankton content of WOD98 was in a raw and basic form. The plankton species were stored without any taxonomic grouping or supplemental indexing. To extract "all copepod data", the user would have to search each and every record for the presence of any one of the over 400 unique copepod taxa contained in the database. The biomass and abundance values were also in raw form, with no quality control, stored in their original as-provided units. To utilize these values, the user would have to use sampling information and metadata to calculate a common unit. Very few users discovered these challenges, however, as the WOD98 data format, designed to efficiently manage five million temperature profiles, made finding and extracting the plankton data nearly impossible. Frustrated users frequently contacted the author directly for help.

World Ocean Database 2001

Over the next three years, more plankton data were added and the short-comings of WOD98 were addressed with the release of World Ocean Database 2001 (WOD01, Fig. 2, O'Brien *et al.*, 2002a).

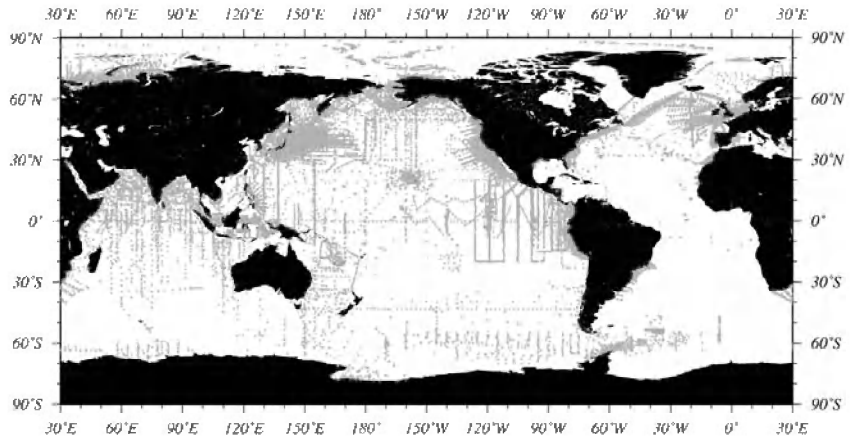

Fig. 2. Plankton tows present in World Ocean Database 2001.

To address taxonomic grouping and indexing, each plankton species was assigned a corresponding Biological Grouping Code (BGC) identification. The BGC worked as a supplemental index which would allow the user to quickly identify and access plankton by major and minor groups (*e.g.*, "zooplankton", "copepods", "chaetognaths", "phytoplankton", "diatoms", "bacterioplankton"). Common units were introduced with the addition of a Common Base-unit Value (CBV) field. Calculated from the sampling metadata, the CBV provided zooplankton in common units of "per cubic meter", and phytoplankton in units of "per liter". Basic quality control was also introduced. Mesh sizes, mouth areas, and towing depths were checked for impossible values. With the addition of the BGC and CBV, it was also now possible to access and examine all values of "phytoplankton" or "copepods" or "diatoms" with automated group-based range checking (*e.g.*, "Is this a reasonable diatom count?", "Is this a reasonable copepod count?").

While WOD01 still used the same data format and layout as WOD98, and thus had the same plankton access problems, the author provided a supplemental online plankton product called World Ocean Database Plankton (WODP). WODP was tailored to the plankton data user, offering additional documentation, content summary graphics, and plankton-specific data files and access software. While this greatly improved access to the plankton data, the access and content were still static. Searching for specific content was still not possible, and the content itself would only be updated every 3-4 years (*e.g.* WOD98, WOD01).

World Ocean Atlas 2001 - Plankton

Shortly after the release of WOD01, global mean fields of zooplankton biomass were created as part of the World Ocean Atlas 2001 series (WOA01, Fig. 3, O'Brien *et al.*, 2002b).

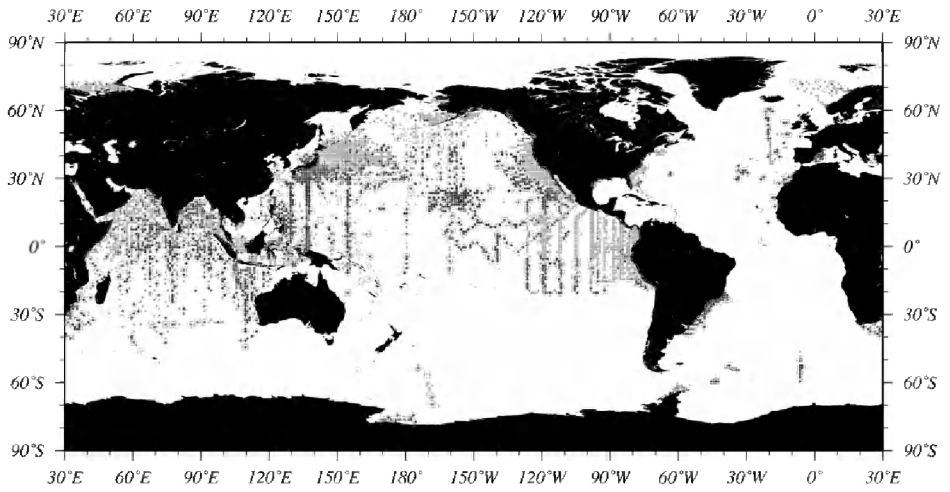

Fig. 3. Mean zooplankton biomass values present in World Ocean Atlas 2001.

During the creation of the WOA01 plankton fields, it became evident that there were still significant data access and usability issues. The creation of these fields represented the first focused effort at actually comparing and combining thousands of plankton measurements from different sampling methods and gear types. The experience not only highlighted the necessity of having complete metadata, but it also demonstrated the necessity of fully understanding and correctly translating the original plankton data and their meaning into any database.

One of the largest metadata problems involved originator-provided taxonomic group sub-total and totals. For example, an investigator reports the presence of "4 apples, 3 oranges, 2 bananas, 9 fruit, 7 vegetables". While obvious to a human reader, if the "9 fruit" is not clearly denoted as the total of all fruit types (*e.g.*, "total fruit"), an analysis program processing millions of fruit records may consider this a fourth fruit category and calculate a total fruit value of 18 from these data. Preventing these types of errors requires careful review of the metadata and data during and after processing and/or digitization. Fixing these errors means reprocessing and/or re-digitizing the mistranslated data.

During the re-processing of these mistranslated data, additional database integrity issues were discovered when comparing the reprocessed data to what was already in the database. These discoveries included lost tows, corrupted values, and disappearing metadata. The causes of these problems ranged from database software errors to limitations within the profile-based database architecture itself. While it was possible to patch and repair many of the problems, the background causes and limitations would always remain a threat to future data integrity. The best solution would be to redesign and rebuild the database.

NMFS-COPEPOD: A New Approach to Plankton Data Management

The Coastal & Oceanic Plankton Ecology, Production & Observation Database (COPEPOD) is a new effort by the National Marine Fisheries Service (NMFS) to provide quality plankton data to the research community. NMFS-COPEPOD is an online database designed specifically for plankton data, developed using the author's 10 years of hands-on experience with plankton data management. In addition to providing a complete re-processing and access to the author's previous content (*e.g.*, WOD98, WOD01), it represents a new focus on user-friendly interfaces, searching, and plankton-specific export formats. Another main focus of NMFS-COPEPOD is to provide clear credit to the associated investigators, projects, and institutes responsible for each and every data set.

NMFS-COPEPOD (http://www.st.nmfs.gov/plankton/) has been online since August 2004. New data are added and available online each month (versus every 3-4 years). As of January 2005, NMFS-COPEPOD contained 86 online data sets, with an additional 40 data sets in final processing and review. Coming in late 2005, new biomass and abundance fields will also be released (Fig. 4, O'Brien, 2005). These will be made available online and in the form of a digital atlas and database product.

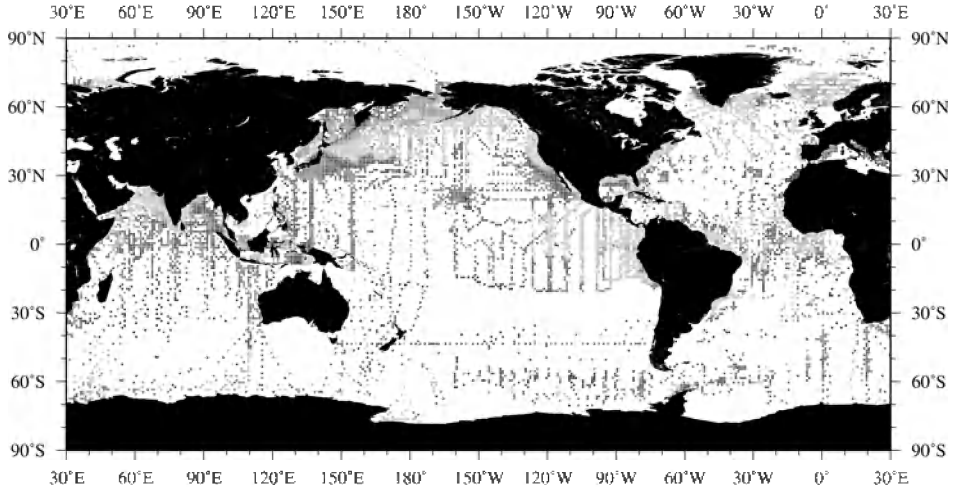

Fig. 4. Mean zooplankton biomass values coming soon in NMFS-COPEPOD 2005.

Conclusions

In the eight years since Hamburg 1996, the metadata requirements for managing plankton data have changed very little. It is the ability to apply these metadata and review the quality of the data that is necessary for managing such data. There is more to building a useable plankton database than just putting plankton data into a database. A plankton data manager needs to know the data (and its quirks and challenges), use the data (applying it and experiencing its short-comings and quality issues), serve the data and the needs of its users (as the reason for building it is ultimately for their use), and

acknowledge the investigators (without whom there would not be any data to manage). A successfully useable plankton database should protect the quality and integrity of its data, serve its community, and thereby encourage submission of future data to the effort.

Acknowledgements

NMFS-COPEPOD is an ongoing effort by the Marine Ecosystems Division of the National Marine Fisheries Service (NMFS) Office of Science & Technology. The content of NMFS-COPEPOD is possible through the efforts and contributions of plankton scientists through the world. Too numerous to list here, the names of associated investigators, projects, and institutes are acknowledged in the "Hall of Fame" section of the NMFS-COPEPOD web site (http://www.st.nmfs.gov/plankton/).

References

Conkright M.E., T.D. O'Brien, L. Stathoplos, C. Stephens, T.P. Boyer, D. Johnson, S. Levitus, R. Gelfeld. 1998. NOAA Atlas NESDIS 25 - World Ocean Database 1998, Volume 8: Temporal Distribution of Station Data Chlorophyll and Plankton Profiles, United States Government Printing Office, Washington, DC. 129p.

O'Brien T.D. 2005. COPEPOD: A Global Plankton Database. U.S. Department of Commerce, NOAA Technical Memorandum NMFS-F/SPO-73, 136p.

O'Brien T.D., M.E. Conkright, T.P. Boyer, J.I. Antonov, O.K. Baranova, H.E. Garcia, R. Gelfeld, D. Johnson, R.A. Locarnini, P.P. Murphy, I. Smolyar, C. Stephens. 2002a. NOAA Atlas NESDIS 48 – World Ocean Database 2001, Volume 7: Temporal Distribution of Chlorophyll and Plankton Data. United States Government Printing Office, Washington, DC. 219p.

O'Brien T.D., M.E. Conkright, T.P. Boyer, C. Stephens, J.I. Antonov, R.A. Locarnini, H.E. Garcia. 2002b. NOAA Atlas NESDIS 53 - World Ocean Atlas 2001, Volume 5: Plankton. United States Government Printing Office. Washington, DC. 89p.