

MARBID: NOAA/NMFS's (US) marine biodiversity database

M.F. Mickevich

National Marine Fisheries Service • Silver Spring, MD USA

Bruce B. Collette

National Museum of Natural History • Washington, DC USA

Introduction

The National Marine Fisheries Service (NMFS) is charged with sustaining and managing the nation's living marine resources. This is done in cooperation with the community at large including members of the regional Fisheries Councils, environmental advocates, politicians, scientists with many varied specialties, fishermen, economists, and policy makers. These people are rarely specialists in biodiversity. The (US) Marine Biodiversity Database (MARBID) is a National Marine Fisheries Service project that filters and rates biodiversity information to facilitate decision making by this diverse group. The ratings are non-parametric Bayesian (relative) confidence limits on the information within MARBID. These are based on metadata, information on data. Individuals using MARBID will have instant

knowledge of the data's taxonomic reliability when making decisions concerning the ocean's sustainability. This paper will introduce the philosophy, components and organization of MARBID.

MARBID contains metadata on its two major components, systematic and distributional data, along with any data that are necessary and not found elsewhere within NMFS databases. Given that its purpose is to follow species of interest to fisheries through space and time, MARBID contains information for species included in the NMFS surveys and has a GIS component. MARBID's distributional information is based on NMFS survey data (Figure 1). Currently, only information and metadata on species from the Gulf of Maine are complete. This includes the fishery's species and other

noncommercial species found within the Gulf's ecosystems and numbers more than one thousand. The systematic data and the metadata are generated within the National Systematics Laboratory.

Taxon Protocol Committee

To develop the metadata standards and consistent protocols for setting Bayesian confidence levels, MARBID relies on the Taxon Protocol Committee. This committee was composed of experts within the fields of computers, taxonomy, and biogeography. Its members included B. Collette, K. Fauchald, T. Munroe, M. Mickevich, M.



Figure 1. The extent of the oceans once surveyed by the National Marine Fisheries Service.

Vecchione, and A. Williams (deceased). This committee established the metadata components and its standards for choosing metadata. It developed protocols based on the characteristics of the data found in the metadata which consistently and objectively rates MARBID's systematic information. The Taxon Protocol Committee has developed protocols to assess the reliability of other components of the database. There are protocols which establish confidence levels for specimen identification. There are standards for deleting or keeping species information within the database after a species name has been changed.

The most serious problem the committee tackled was that of organism names. Names are the means by which collected specimens are identified. Evolutionary trees are the tools of scientific work. Both functions are necessary. The direct results of the science of systematics are two principal products, classifications of names and cladograms which organize these named groups into an evolutionary tree. Only a few systematic studies perform both functions. Both use the same Linnean names such as a genus, species, family, phylum, etc. These functions are not the same. First, the naming of organisms and their grouping into Linnean categories do not necessarily produce the same species groups as implied by an evolutionary tree. Classifications are less informative than phylogenetic trees. Classifications can also contradict trees. This is especially true for marine invertebrates. Second the history of naming follows the rules for nomenclature and the association of these names with type specimens. Effort is made to keep the first name associated with the type a constant, all other specimens from other localities may be assigned any other name. The variation found within organisms is mapped by type assignments. These types are deposited into museum collections. The naming of these types is conserved as much as possible. This activity is independent of forming an evolutionary tree. Most studies obtaining species trees do not consider types or type localities, yet these identify specimens by names. If the taxonomy of a group is changed by a taxonomist after such a study, the evolutionary tree can be meaningless.

Although scientific standards of a systematic study may be high, the information generated may have low reliability. In addition to the problems mentioned above, all names are not equally reliable. Some species names are well established scientifically while others are recognized as tentative. Species, especially of marine invertebrates described in the nineteenth century without benefit of modern methods, may not have been restudied and reconfirmed. These implications are straightforward: some species names as used, may not correctly represent true species. These may be part of another species, may belong to multiple species or may be composed of two or more unrelated groups. An

uncritical database may, therefore, convey misleading information. MARBID uses protocols on metadata to establish confidence limits for all taxa.

Systematic Confidence Ratings

MARBID's aim is to express scientific complexity visually. The classification is listed and the species tree is graphed. The NMFS database conveys the reliability of a species name readily to all users with visual cues, i.e., colors: green for highly reliable, yellow for less reliable, and red for poorly established groups. A species could be green whereas the genus it belongs to might be yellow. MARBID reflects the reliability of taxonomic or of species' tree groups. The graph of a species tree might have some branches colored yellow, red, or green, depending on the underlying reliability. MARBID's Geographical Information System maps species' geographic distributions in color to reflect that species' reliability.

To rate taxa in such a fashion the Taxon Protocol Committee has established protocols that together form an expert system. These protocols are used to screen scientific publications on the species being considered for necessary characteristics. For example, were the types of the nominal species examined? Were the classifications and species trees produced by an objective method? These protocols include two major categories: Comprehensiveness and Quality of Analysis.

Comprehensiveness includes the following criteria: materials, types and taxonomic characters. Materials deal with the organisms under study: How complete was the examined material? Did it include the original type specimens? Did the material examined represent the total geographic range? Were all the species within the group examined? Types involve the questions: Were the new specimens identified properly? Were types for new species deposited in archival museums? Characters deal with how much information examined on each specimen. Was it complete and sufficient?

Quality of Analysis contains four criteria involving the choice of the species group, and the analysis of the data. The first criterion involves characteristics concerning the choice of the species being studied. How reliable are the species being studied? The second, concerns the objectivity of the quantitative analysis performed on the species' data. The third, the justification for placing the species group within a generic or familial level. The fourth requires that there is at least one character held in common by all and only those species included within a group. This is called a diagnostic character.

These criteria are used within protocols to code a species red, yellow or green. As an example, the characteristic, Types has two criteria: 1) Newly collected material were identified according to standard; 2) New species' type material has been deposited. If a system-

*MARBID's aim is to
express scientific complexity
visually.*

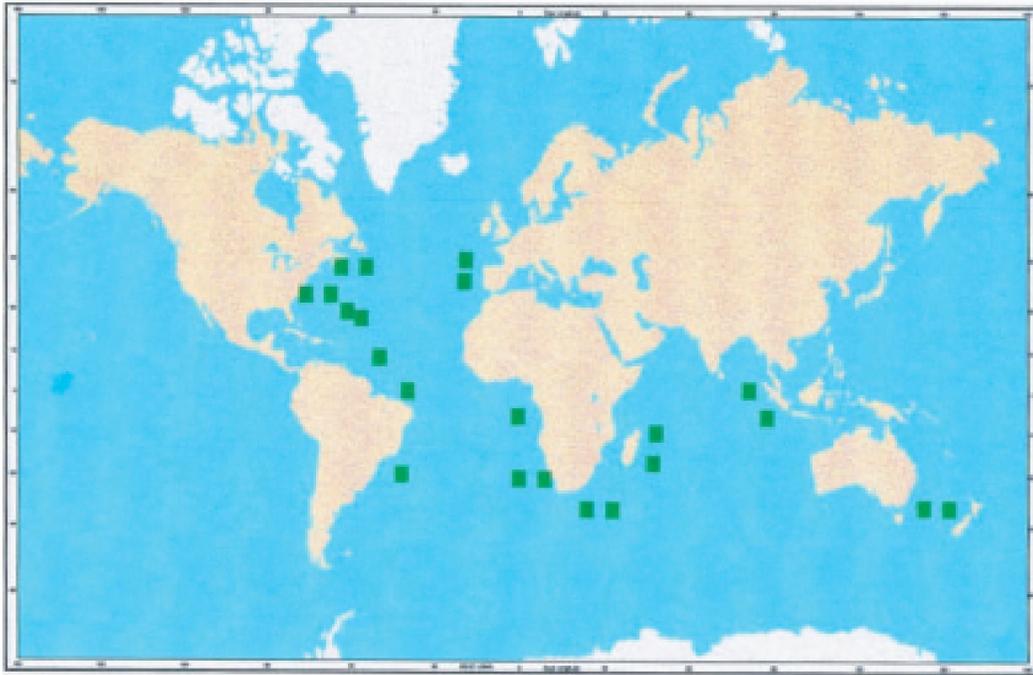


Figure 2. The locality information of a green species mapped onto the oceans.

atic study has both characteristics it is GREEN, if the second is missing it is RED, if only the second is present, the study is YELLOW. Each of the characteristics for each category is examined. If all are green, the group is green. If one is red, the group is red. All else is yellow.

Geographic Component

MARBID's geographic component is based on NMFS survey data (Figure 1). NMFS survey data has a history exceeding a century. Managing the nation's living marine resources requires population information for all commercial fish species and their associates through fine time slices. Survey data collected at regular intervals provide this information. NMFS survey data are as rigorous a sampling as can be found within the scientific community. For example, the Northeast Fisheries Science Center's (NEFSC) surveys cover the New England and middle Atlantic regions including both Georges Bank and the Nantucket Shoals (Azarovitz, 1981). The gear and collecting regime are constant. Annual surveys have included a fall and spring ground-fish survey for the last thirty years. There are an annual scallop survey and a shrimp cruise. A winter ground-fish survey has taken place for the last three years. Recently, a summer Gulf of Maine survey has been added. In addition, there is a triennial shellfish cruise. These sampling regimes are developed according to a statistical model that reflects the distribution of the species considered. This model uses a stratified random sampling design that produces an estimate of the species inventory found within that region (Grosslein, 1969; Anonymous, 1988). Not all species are identified

and inventoried (Flescher, 1980) because of personnel and time constraints, but more could be identified if it were deemed of high priority. Names of the identified species, the biological variables, and the physical parameters are stored within Oracle databases according to the station. Stations are identified by longitude, latitude, depth, gear, date, and time. The station data along with the identified species are used within MARBID as distributional information and mapped within its GIS component.

MARBID utilizes locality information on species monitored by NMFS surveys (Figure 1). Locality data are organized by collecting stations. Species occurrences are recorded by their codes and referred to by common names. These codes are characteristics of regional surveys or regional laboratories. The same number can indicate different taxa in different regions. The common names are associated with scientific names but these scientific names may not be those that reside in MARBID. The species codes and names are altered to fit into MARBID.

Other Components

MARBID follows changes both in the evolutionary relationships and within the Linnean classification systems. These changes can have profound effects on the observations taken on species. For example, changes involving species split into two or more species which are not closely related have the following result: the prior locality data on the original species is VOID. Changes which result in splitting one species into two or more species with intermixed geographical distribu-

tions have the following result: the observations can be used at the generic level but not at the specific level. There are multiple rules by which locality information must be moved to other taxa, voided or reassigned. MARBID manages changes in observation assignments where these are due to changes in classifications and/or phylogeny of taxa.

There are other components to MARBID, a Geographic Information System (GIS) component and a Locality History File, to name a few. All have rules and change tracks for database management purposes. Each is necessary for following species through time and space. The GIS component performs the mapping. The Locality History File holds station data which has been reassigned or voided.

Following species through time and space for fisheries management and marine conservation is not MARBID's only utility. If a species' distribution is mapped on the globe with the fishery's survey information, the localities can be colored reflecting that species' taxonomic reliability (Figure 2). Green would indicate the distribution of a very reliable species, red, a not so reliable species. Generic distributions could have red, green and yellow locality indicators. When a species being studied within an ecosystem is red, the question of whether the species is correctly defined and identified has bearing on the scientific results. When, in the analysis of ecosystems some or most of the species present are rated red, others rated yellow, and there are very few greens, the results of such a study are not as meaningful as one employing only green species. Conversely studies on entirely green species can be regarded as more definitive.

Discussion and Conclusions

MARBID is designed to accommodate as many observational and data problems as possible in order to be effective. In order to do so, it had to move away from present designs in biodiversity databases which are taxonomic or phylogenetic by nature. Museum collection databases are composed mostly of elements used for internal bookkeeping. The major costs are associated with considerations of maintaining and gathering this information. Taxonomic databases are lists of names that are useful in themselves or are necessary for nomenclature studies. Nomenclature studies follow

name changes. The phylogenetic systems presently implemented, present classifications as species trees and do not consider the quality of the conveyed information or the fact that trees can be different depending on the characters and methods used. Use of one of these 'species trees' may be highly misleading and lead to erroneous conclusions in a biodiversity study. When a classification is not the species tree and vice versa, the database should so inform the user.

Many alternative classifications exist for a group of species, at least for the common fishery's species. The evidence for one classification over any alternative should be included. Systematic databases should be structured or designed to reflect these

characteristics of systematics.

MARBID exists to help us manage the nation's living marine resources. The nuances of systematic and biogeographic information are obvious to trained practitioners. As stated above, the members of the community at large, are involved in decisions of critical importance to ocean sustainability. They are not specialists. The increased amount of time necessary to gather the metadata in order to rate and filter all the locality and systematic data within this information system is justified on this basis alone. Biodiversity databases which only present data which is not filtered or rated with constant, rigorous, and objective protocols can only be looked upon as questionable.

LITERATURE CITED

- Anonymous, 1988: An evaluation of the bottom trawl survey program of the Northeast Fisheries Center. *NOAA tech. Mem.*, NMFS-F/NEC-52, 1-83.
- Azarovitz, T.R., 1981: A brief historical review of the Woods Hole Laboratory trawl survey time series, p. 62-67. In: *Bottom trawl surveys*. Doubleday, W.G. and D. Rivard, eds. *Ca. Spec. Publ. Fish. Aquat. Sci.*, 58, 273pp.
- Carpenter, K.E., B.B. Collette and J.L. Russo, 1995: Unstable and stable classifications of scombroid fishes. *Bull. of Mar. Sci.*, 56(2), 379-405.
- Flescher, D.D., 1980: Guide to some trawl-caught marine fishes from Maine to Cape Hatteras, North Carolina. *NOAA Tech. Rpt. NMFS Cir.*, 431, 1-34.
- Grosslein, M.D., 1969: Groundfish survey program of BCF Woods Hole. *Comm. Fish. Rev.*, 31(8-9), 22-30. 

*MARBID exists
to help us manage
the nation's living
marine resources.*