CephBase: testing ideas for cephalopod and other species-level databases.

James W. Wood
Univ. of Texas Medical Branch • Galveston, Texas USA

Catriona L. Day
Dalhousie University • Halifax, Canada

Phillip Lee
Univ. of Texas Medical Branch • Galveston, Texas USA

Ronald K. O'Dor
Dalhousie University • Halifax Canada

Introduction

This report shares our experience of what worked and what did not while developing CephBase (www.cephbase.dal.ca). We share our triumphs and failures as well as our vision for the future to help and encourage those starting new database projects as well as outline ideas about collaboration between existing and yet to be developed databases. It also explains why we chose cephalopods and urges readers to share their enthusiasm for the groups of marine animals they investigate. Clearly, new Internet technologies offer a chance for substantial improvement in global distribution and sharing of information.

What is CephBase?

CephBase is a dynamic html (dhtml) relational database-driven interactive web site. The current prototype version of CephBase was developed at Dalhousie University in Halifax, Canada and was sponsored by the Sloan Foundation following the Workshop on Non-Fish Nekton in Boston, December, 1997. The goal of CephBase is to provide a globally accessible platform for information on taxonomy, life history, distribution, fisheries, and ecology for all living cephalopod species (i.e. octopus, squid, cuttlefish and nautilus). We have begun to accomplish this goal but have barely scratched the surface of what could, and should, be done. The pilot project was also designed to help define the goals of the Census of Marine Life (CoML) by demonstrating the possibilities of dynamic web sites. The web is an ideal medium for collaboration and communication of species-level data. Interest in this new tool is rapidly growing.

CephBase is part of the CoML project funded initially by the Sloan Foundation. The Census of Marine Life seeks to "assess and explain the diversity, abundance, and distribution of marine life" (Ausubel, 1999). The CoML will accomplish this by making existing data more accessible and useful to researchers, and by creating databases that can accommodate new data generated by the application of new technologies in order to find and identify marine species (Malakoff, 2000). CoML recognizes that our knowledge of species that live in the ocean lags behind our knowledge of species that live on land. It also recognizes that many stocks of fish have been critically over-fished and that there are thousands of new marine species awaiting discovery.

Why Cephalopods?

Cephalopods are an ideal test group as there are only 700 known species in the class. This makes the database very manageable for the development of new techniques. They are also invertebrates, which are typically understudied despite the fact that invertebrates make up 95% of the animal kingdom. Cephalopods are active, have large brains and well-developed sense organs (Hanlon and Messenger, 1996). They are capable of vertebrate-like behavior, as well as rapid color, shape and texture changes unrivaled in any other animal. Perhaps it is these features that make them popular with the public.

Cephalopods have short life spans, fast growth rates (exponential when young), and they tend towards semelparity. They are quite distinct from fish, not only

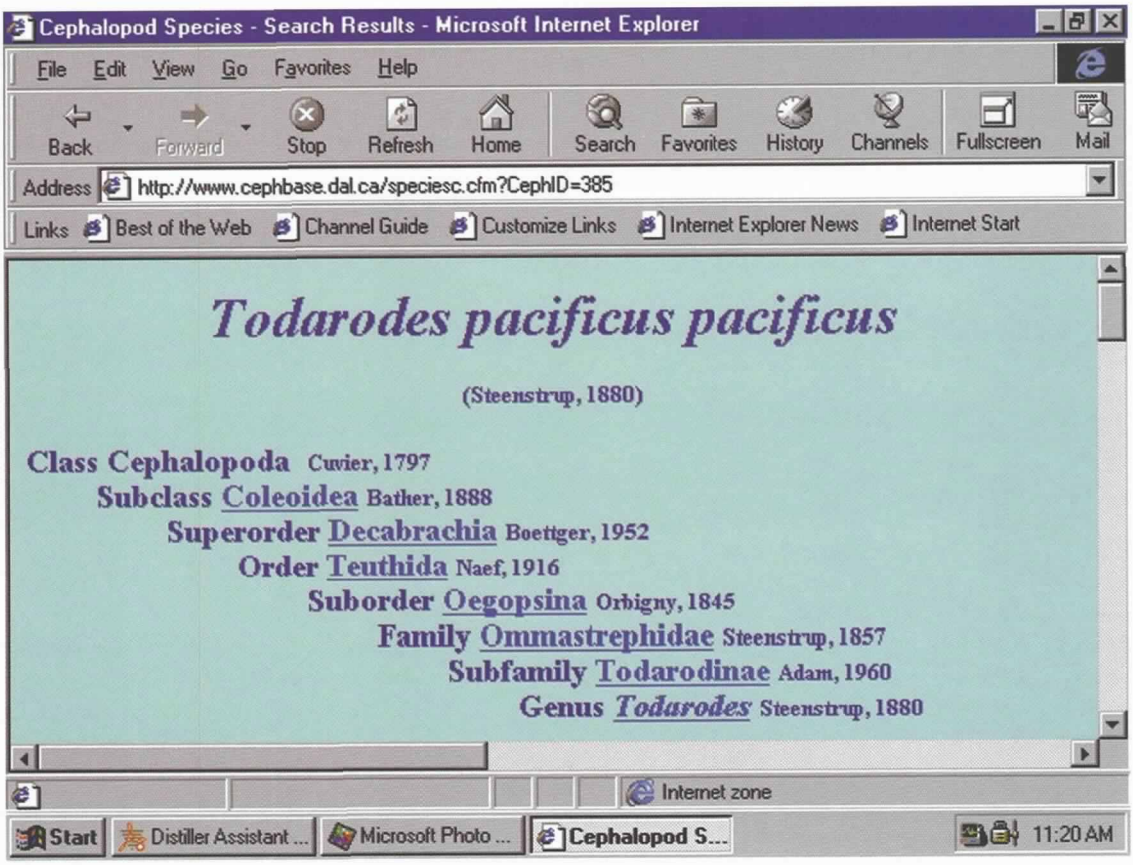

Figure 1. A species page from CephBase detailing the taxonomy, synonymy and common names of a commercially important squid, the Japanese flying squid (Todarodes pacificus pacificus). *The information for prey and predators and specimens cannot be seen in this view but are accessible by scrolling down the page.*

in their morphology but also in their life history (O'Dor, 1998). They are all predators but never top predators. Despite the low number of species, cephalopod fisheries are a significant and growing percentage of the world's catch. In fact, cephalopod biomass has been estimated to equal that of all the species of fishes in the world's oceans (Clarke, 1987). CephBase lists many of the species of fish, birds, and marine mammals that prey on cephalopods. Models developed for fish have not been able to adequately estimate cephalopod life cycles. For example, using length frequency data (ELEFAN) to model the life span of the smallest species of cephalopod, *Idiosepius pygmaeus*, yields a life span of 3 to 10 years; much more than their actual maximum life span of 80 days (Jackson and Choat, 1992)! There is a large potential for selective fishing and temperature variation to greatly influence the population structure of short lived and often semelparous cephalopods such as squid (Forsythe, 1993; Rodhouse et al., 1998). From a fisheries management perspective, much more biological information is needed; we hope to be one of the key information providers.

. . . our goals are to provide species-level data on all cephalopods on a reliable and globally accessible platform.

CephBase was developed independently but is similar in concept to FishBase (www.fishbase.org) and is global in species coverage. All 703 recent species and subspecies of cephalopods are covered following Sweeney and Roper (1998). Cephalopods live from the tropics to the poles, from the intertidal zone to the abyss, from planktonic to benthic; thus CephBase covers species that live in all typical marine habitats. CephBase provides contact information for almost all of the world's cephalopod specialists; part of our role is to encourage data sharing and collaboration.

In summary, our goals are to provide species-level data on all cephalopods on a reliable and globally accessible platform. Such a global database will facilitate collaboration both within the cephalopod community and among all marine sciences.

How does CephBase work?

Under the hood of CephBase you will find a Microsoft Access database. This relational database holds the data in various tables. Instead of one big table, a relational database holds many smaller tables that are linked, or related, to each other. This makes accessing

the database much faster for users and conserves disk space. SQL (Structured Query Language) is used to manipulate these data and a Windows NT server linked with Cold Fusion (an API or Application Programming Interface) serves dynamic web pages to users. This allows our database to be fully accessed by anyone with a computer linked to the Internet.

One of the major benefits of this structure is that as soon as new data are added to the database, all web pages that use that data are instantly updated. We will say more about this below when we discuss the advantages of generating real-time distribution maps.

What are the current features of CephBase?

1. *Classification of all known cephalopods*
   We currently have all the taxa, authorities, and the year the taxa were described on-line for all of the 703 known living species of cephalopods listed by Sweeney and Roper (1998). Information on a particular species can be quickly located by using the search engine; results are listed in table format. Users simply click on a species in the table and all the taxonomy, from class to subspecies, for that cephalopod is displayed. Users can also display an alphabetized list of all cephalopod genera. Clicking on a genus leads to a list of all species it contains. For each species, synonymies, type repositories, type localities, references and common names are listed (Figure 1). References are listed in abbreviated form and access to full references is only a click away. This demonstrates what can be done and it is not hard to imagine life history, ecological and morphological data being accessed in the same manner.

2. *Distribution maps*
   To answer the question, "Where does it live?" we currently have over 3,150 referenced localities for about 320 species in our database. Maps are generated in real-time using the Xerox Parc Map Viewer (www.parc.xerox.com/parc-go.html). All latitude and longitude data used to generate maps are from published sources and are listed in tables and referenced. In many cases, the individual specimens used to populate the database can be tracked to a museum repository. While there are occasional problems with the unsupported and free Xerox Parc server used, we definitely recommend this approach. When a user clicks on a species distribution map he or she can zoom in from a global view to any local area in the world to view details (Figure 2). This option alone would take thousands of static maps for each species which would take up a lot of hard drive space. Furthermore, when we add a new record to the database, it is automatically and instantly included in maps requested by users. The alternative is to plot all those points by hand on scores of static maps for each species, then upload them every time new data are added. Generating maps in real-time from a database is the best way to map specimen locations on-line and was first done for a marine species by CephBase. This method empowers users by letting them select the information that they would like displayed with the options they want.

3. *Ecological data*
   Ecological data are needed to fit cephalopods into global models. At present, CephBase holds 420 predator records and 979 prey records. All are referenced and all have the complete Latin name of the cephalopod involved in the interaction. Papers that just list "squid" or "cephalopod" are not included.

4. *Images of selected species*
   The Internet is a graphical medium so we try to provide on-line images of as many species as possible. The illustrations of species are a necessary tool to support taxonomy and ecology of cephalopods as well as a tool for educators. At present, links to pictures of 30 selected species are available from The Cephalopod Page (http://is.dal.ca/~ceph/TCP/index.html) maintained by one of the authors, J. Wood.

5. *Directory*
   We maintain the International Directory of Cephalopod Workers to help foster global collaboration. In addition, a number of scientifically related links have been established and maintained, allowing someone new to web searches or the world of cephalopods to find other cephalopod resources quickly.

What are the future directions for CephBase?

The CephBase team (NRCC, Dalhousie University and National Museum of Natural History of the Smithsonian Institution) has recently been awarded a National Ocean Partnership Program (NOPP) grant (#2000-5-16 AP) to greatly expand the database. The NOPP is composed of 12 federal agencies and the Sloan Foundation. The Sloan Foundation, Office of Naval Research and National Science Foundation have joined to fund the Ocean Biogeographic Information System (OBIS), a major initiative directed toward the development of the Census of Marine Life. Some of our objectives for taking CephBase to the next level are outlined below. Because many of these objectives are similar to those of other species-level projects, they offer a glimpse into the near future and provide a starting point for future standardization.

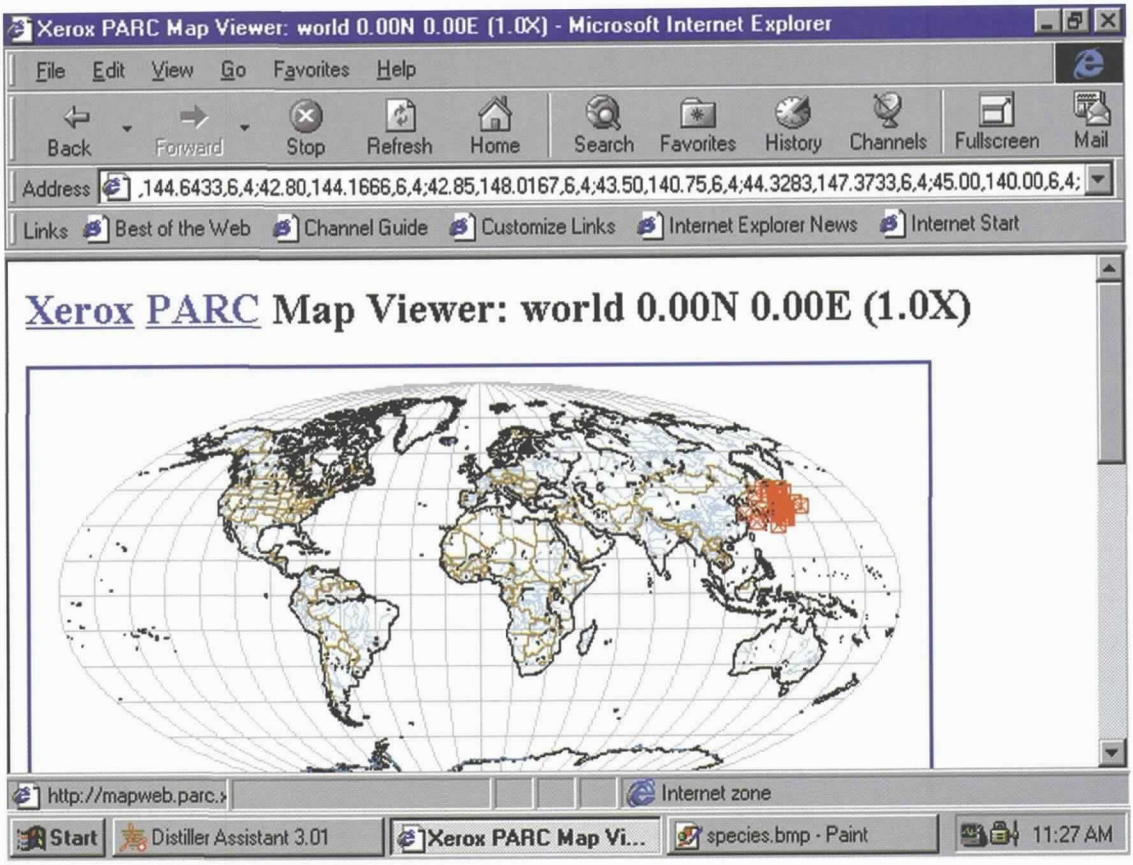

Figure 2. A distribution map for specimens of Todarodes pacificus pacificus *accessed from the species page of CephBase linked with the Xerox Parc mapping system. This map can be manipulated by zooming in and out by clicking on different parts of the world.*

1. *Improved real-time species mapping*
CephBase has been generating maps in real-time for over two years. FishBase and other web sites have not yet taken advantage of this dynamic technology. As mentioned above, we feel that generating maps in real-time from the database is the method of choice for a constantly growing species-level database. It empowers users to search for and display the information they want in the form they need from a simple web page. Additionally, by generating maps in real-time, users are not required to purchase expensive software or download large files. The complexity of the database and dynamic web site code will continue to remain completely transparent. A fifth grader will be able to access the data but it will be most useful to persons interested in the global biodiversity of this ecologically and economically important class (e.g., fisheries scientists, taxonomists and government resource managers and policy makers). We plan to move away from Xerox Parc to a more powerful and dependable map server where we can overlay environmental conditions (currents, depth, temperature, bottom substrate, etc.) on distribution maps as well as color code specimens to indicate their maturity, sex and depth of capture. Ideally users would be able to click on any section of ocean and have access to the records of all species of cephalopods in that area. It is our plan to collaborate with OBIS by providing test data to develop this and other tools for all of the CoML projects.

A fifth grader will be able to access the data but it will be most useful to persons interested in the global biodiversity of this ecologically and economically important class . . .

2. *Add life history, physiology and fisheries fields to the database*
The taxonomy for cephalopods is already on-line. While very useful, these data were the easiest to collect (most of them were in one recent, authoritative revision) and integrate into our database as their organized hierarchical form suits relational databases extremely well. This taxonomic work is simply a framework that can support a number of other types of species-level data. Life history traits such as age at maturity, size at maturity, hatchling size, fecundity, maximum and minimum size at

maturity as well as physiology such as temperature, growth and metabolic rates should be added. This kind of information is often hard to access as it is spread out in many technical journals. Currently we are exploring options through the Cephalopod International Advisory Committee (CIAC) to expand the fisheries content of CephBase. We will also approach the Food and Agriculture Organization (FAO) and appropriate Asian partners (Asian Fisheries Society).

3. *Continue to populate the database with predator, prey and location data*
Cephalopods are important predators but never top predators as they in turn are preyed upon by fish, birds and marine mammals. What cephalopods eat, and what eats them, links them to many other animals in the ocean including important commercial species of fish and high profile groups such as whales. Cephalopods have short life spans typically of a year and in that time they rapidly move up the trophic levels (Levi et al., 1999). Like all data in CephBase, the predator and prey data set is fully referenced and therefore helps researchers who are not familiar with cephalopod literature quickly find the information they need. We hope to eventually interface with other databases in order to begin on-line reconstruction of the complex predator and prey interactions of marine animals.

   The CephBase team has started to add predator and prey data from sources found while working on other projects. While we already have an impressive data set, we will now make a directed effort to actively search for this information.

4. *Take full advantage of images and video available at NRCC and Smithsonian*
The World Wide Web, as most people know, is graphics intensive; so are cephalopods. Cephalopods can change texture, body shape and color in the blink of an eye. Color images and video clips are rich in useful information with high impact with site visitors but are difficult or impossible to cost effectively incorporate into traditional printed media such as scientific journals. Image data are especially important for cephalopods as species can be identified based on their changing body patterns and cephalopods may communicate with visual signals (Hanlon and Messenger, 1996). These data are lost upon preservation.

   Despite this, the current version of CephBase does not have any graphics of its own except the banner image on the main page. The NRCC houses many live cephalopods and the scientists at NRCC have been all over the world for cephalopod research. They have over 5,000 35 mm slides and an extensive collection of video footage. Many of these images will be merged with the database to take full advantage of the technology, including live footage of rare species as well as an overview of the unique behaviors of cephalopods.

5. *Increase collaboration*
To expand, CephBase will require more data. We are excited about this new technology and think that a central hub to store cephalopod data makes sense. We plan to develop a collaborator database to list all those that have donated their data to this project. We intend to involve cephalopod researchers from every continent (e.g. CIAC). This solution will eliminate duplication of effort and will reduce the time users search for references. This solution does not have to compete with regional projects; rather, it gives scientists an additional outlet for their data that is very powerful and rapidly growing.

6. *Generate positive publicity and education about marine life, Census of Marine Life project and the NRCC*
The wealth of scientific knowledge, while technically available to the public, has traditionally been held in specialized journals housed in a limited number of large libraries. The public infrequently accesses these institutions. Most people simply do not know that they can access this information, or they do not have the skills to find and understand articles of interest, or they live too far away. Eventually some of this information is republished in popular articles and in textbooks. Only a small fraction ever becomes available to the layperson and this information may take years to decades before it becomes available. However, we have found that the public, from grade school kids, to divers, aquarium hobbyists, and educators are very interested in cephalopods.

   On-line dynamic species-level databases are a relatively new medium with new properties. One of them is that scientists can rapidly and directly publish quality data that are easily accessed by a potentially large audience. On their web sites, scientists can control the quality of this media and are not forced to "dumb down" content for the lowest common denominator.

   There are many individuals besides our scientific colleagues who are interested in the work we do (e.g. fishermen, policy makers, environment groups, hobbyists, teachers and students) and the animals we study. Web sites are an excellent tool for giving something directly back to the public who fund research through tax dollars. Concepts like the Census of Marine Life should be fully developed to grab the public's attention and focus it on an understudied and fascinating part of the planet, the oceans.

In addition to having all of the database available on-line, the CephBase team plans to develop several areas specifically for the general public and for educators. The initiatives outlined below will bring in a much broader range of users that the pilot version of CephBase did not target:

- An on-line key for species identification
- Virtual dissection guides
- An on-line visual cephalopod museum
- An explanation of cephalopod color change with supporting video and animations
- A demonstration of how cephalopods use ink decoys with supporting graphics
- An explanation of their jet propulsion with supporting video and animations
- A webcam on octopus or cuttlefish tanks to attract younger students
- A fun interactive test "So you think you know cephalopods?" for students

7. *Move to client-server architecture for increased security*
Currently CephBase, like most other databases running on Windows operating systems, is using Microsoft Access, a shared file database. Our "international directory of cephalopod workers" was originally designed to be updateable from the web. This worked well for over a year as scientists who moved entered their own new addresses, phone numbers and email address, saving a lot of time. However, after a while, a few people stared putting in garbage data and even worse, large chunks of data would disappear. Although it is hard to believe, shared-file databases like Microsoft Access are susceptible to loss of data caused intentionally or unintentionally by users, data corruption, and they are much slower (Forta, 1998). To eliminate this problem, one of the first things we will do is move CephBase to a more secure client-server database such as MS SQL Server 7.

8. *Ensure that all data are fully accessible on-line to users and that access does not require any special software skills*
The most powerful and complex software is useless if not utilized. We will continue the precedent of ensuring that the complexity under the hood of the database driven web site is transparent to users and the site is fully accessible by standard browsers such as Netscape and Explorer. Dependence on special downloadable files, programs, high end applications that only work on some browsers, special viewers or odd file extensions will be avoided unless absolutely necessary. Our clients are biologists, educators, fishermen, government employees and students, not computer science majors!

Our clients are biologists, educators, fishermen, government employees and students, not computer science majors!

Additional Ideas for On-line Biological Databases

We feel very strongly that a generic blank species-level database should be created and supported. This would be of incredible worth for new projects on taxa not yet covered for three key reasons: 1) new projects that all start from the same point and use the same protocols will be a lot easier to integrate with each other in the future; 2) it will be easier to troubleshoot and develop advanced features for an array of similarly set-up databases using the same software than for databases using different software; and 3) a blank "plug-and-play" species-level database would encourage others to start new database projects on groups in their area of expertise. This idea would need to be well supported to work. However, we feel that it would pay high dividends by jump starting new database development and facilitating collaboration among developers.

Another idea is to help promote each other's sites by creating a more organized family of marine species-level databases. Members would provide links to all other members. Ideally these links would be placed in the same location on each site so users would have some consistency between sites. Minimal standards should be met before a database is included.

It would also be useful to create a public list-server on the development and administration of biological relational databases. Many of the site developers, like the authors of this paper, are biologists and experts in their field but lack formal computer science training. Developers of new sites and those expanding older ones are all likely experiencing many of the same problems. Some mutual help and teamwork would likely pay large dividends.

A set of standards for on-line species-level databases is needed. The key is for this initial group of experts (biologists, taxonomists and computer programmers) to formulate a win-win plan for encouraging collaboration from all specialists that study a taxon. As databases grow and additional collaborators come onboard, quality control will need to be continually assessed. Garbage data or bad coding can cause incredible amounts of work. Avoiding these problems as we grow will be a priority.

Conclusion

In retrospect, this article is more about solutions to old questions than about computers, GIS software and databases. Questions like: Where should I publish this growing mass of data so that they will benefit the most people? How can I use them to examine the world in a new way? Where do different species live in the ocean and how do they interact with each other and their

environment? Technology is not the question, but it is a growing part of the answer.

REFERENCES

Ausubel, J.H, 1999: The Census of the Fishes: An Update. http://core.cast.msstate.edu/censFeb99.html

Clarke, M.R., 1987: Cephalopod biomass-estimation from predation. In: *Cephalopod Life Cycles Vol. II Comparative Reviews.* P.R. Boyle, ed., Academic Press, London, 221-238.

Forsythe, J.W., 1993: A working hypothesis of how seasonal temperature change may impact the field growth of young cephalopods. In: *Recent advances in cephalopod fisheries biology.* T. Okutani, R.K. O'Dor and T. Kubodera, eds., Tokai University Press, Tokyo, 133-144.

Forta, B., 1998: *Cold Fusion Web Application Construction Kit*, 2nd ed. Que Corporation, Indianapolis, IN, 1024.

Hanlon, R.T. and J.B. Messenger, 1996: *Cephalopod Behaviour.* Cambridge University Press, UK, 232 pp.

Jackson, G.D. and J.H. Choat, 1992: Growth in Tropical Cephalopods: An analysis based on statolith microstructure. *Can. J. Fish. Aquat. Sci.*, 49, 218-228.

Levi, C., G. Stone, J.R. Schubel, 1999: Censusing non-fish nekton. *Oceanography*, 12(3), 15-18.

O'Dor, R.K., 1998: Squid life-history strategies. In: *Squid recruitment dynamics, FAO Fisheries Technical Paper. No. 376.* P.G. Rodhouse, E.G. Dawe and R.K. O'Dor, eds., Rome, FAO, 233-254.

Malakoff, D., 2000: Grants kick off ambitious count of all ocean life. *Science*, 288, 1575-1576.

Rodhouse, P.G., E.J. Murphy and M.L. Coelho, 1998: Impact on fishing on life histories. In: Squid recruitment dynamics. *FAO Fisheries Technical Paper. No. 376.* P.G. Rodhouse, E.G. Dawe and R.K. O'Dor, eds., Rome, FAO, 255-268.

Sweeney, M.J. and C.F.E. Roper, 1998: Classification, type localities and type repositories of recent Cephalopoda. In: *Systematics and biogeography of cephalopods*, vol 2. N.A. Voss, M. Vecchione, R. Toll and M.J. Sweeney, eds., Smithsonian Institutions Press, Washington, 561- 582.

OCEANOGRAPHY SOCIETY · 1988