Where are all the data?

Gordon Paterson, Geoff Boxshall, Neil Thomson and Charles Hussey
The Natural History Museum • London UK

Introduction

In the summer of 1869 H.M.S. *Lightning* undertook a cruise to investigate the deep waters to the west and north of the British Isles. Thus deep-water oceanography was launched. Since that time biological exploration of the world's oceans has gone through three phases (Wüst, 1964): the age of exploration, where nations sent out vessels to travel the world; the era of institutes, where research became focussed on the activities of large oceanographic institutes and the research vessels they ran; and, currently, the age of research programs, driven by groups of researchers and funded from international as well as national sources. In parallel with deep-water oceanography, there has been extensive research on coastal waters and on fisheries. When taken as a whole, the potential available data represent an enormous investment and significant resource. There are considerable advantages in bringing such data together for programs such as the Census of Marine Life and the Ocean Biogeographic Information System. In this article we examine where such datasets are to be found, the relative advantages of using existing data, their limitations and how such information could be made available.

. . . merely relying on literature will miss a considerable body of information.

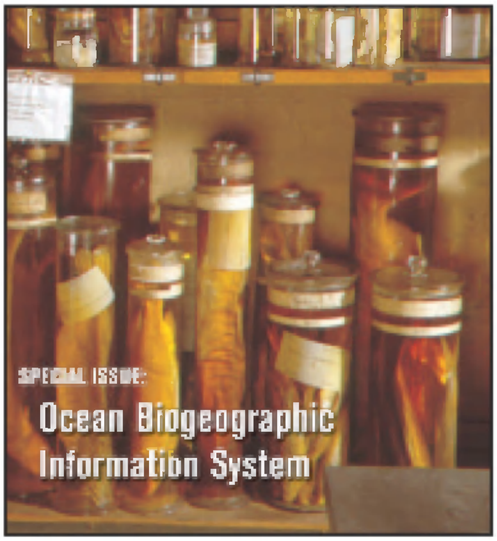

Figure 1: Biological data off the shelf. Museum specimens, like these fish, represent a considerable source of taxonomically verified information. (Photo courtesy Paul Clark Natural History Museum)

Existing data

Marine biologists are not alone in initiating databases that mobilize disparate datasets. Considerable advances are being made by researchers seeking out information from disparate sources. Advances in our understanding of the physical and chemical environment of the world ocean and in climatology have come from analysis of existing chemical, temperature and oxygen records held in the World Oceanographic Data Centers (Levitus, 1996; Conkright and Levitus, 1996). In biological research an assessment of the biodiversity of Guyana was achieved by pooling existing knowledge about birds, amphibians, mammals, butterflies, and various plant groups (Funk et al., 1999). Similarly ICLARM have spent considerable effort to bring together a comprehensive database of fishes – Fishbase (http://ibs.uel.ac.uk/fishbase). Such studies have demonstrated that there are pragmatic advantages to collating existing data. Collecting new material will be costly and ultimately provide a limited number of datapoints. Also new expeditions will require focused scientific rationale to gain funding. A greater understanding of where there are major gaps in our understanding of deep-ocean systems could be derived from analysis of existing data, thus providing a valuable focus for future research.

Sources of Existing Data

Most of the major findings from oceanographic research, particularly from the earlier phases, can be found in the scientific literature. Electronically capturing this information would provide a useful way to populate quickly systems such as OBIS. In the cases of fisheries records this is the only way such data can be amalgamated. Yet merely relying on literature will miss a considerable body of information. There are several places where specimen-based data from the past century are stored, some of which have never been published in the literature: 1) museums; 2) oceanographic institutions; 3) universities; and 4) commercial companies.

1. ***Museums and Collection Storage Facilities.*** These institutions are an enormous source of information and of associated taxonomic expertise (Figure 1). Many museums already have

much data electronically stored, particularly those facilities based in the U.S. European museums are currently carrying out databasing initiatives. For example, within the NHM, specimen information is available electronically from 1994 and there are currently a series of projects to integrate other records, particularly type specimen information. Museum collections could provide a vast coverage of both spatial and temporal information, across many taxonomic groups.

2. *Oceanographic Institutions*
   Many institutions hold records of species abundance and location, although they probably have limited specimens collections. For example in the UK, there is a large pelagic organism database maintained at the Southampton Oceanographic Centre. Similarly, in France at IFREMER there is BioOcean – an Oracle database of biological information from various French deep-sea studies. Most of the records are derived from the identification of specimens by taxonomic specialists. The Alfred Wegener Institute has databases on Polar Regions.

   From the 1950s through the following two decades oceanographic research by institutes in the former Soviet Union generated enormous amounts of specimen data from all over the world ocean. It is likely that records of these collections will not be available electronically, although there are important specimen records in institutes such as the PP Shirov Oceanology Institute in Moscow. While Oceanographic Data Centers manage physical and chemical data, several have biological information which could contribute to OBIS (see references in Proceedings of the International Workshop on Oceanographic biological and Chemical Data Management, 1997).

3. *Universities (including Marine Biology Field Stations)*
   Universities and coastal marine stations are a source of abundant local or regional species data, often stretching over a considerable time span. Much of this information is available only in manuscript, in the form of theses and restricted circulation publications; few original collections remain, apart from those used in teaching. Potentially, this category could be a rich source of information with which to develop long-term change models.

4. *Commercial Companies and Environmental Monitoring Programs*
   All across the globe environmental impact and monitoring programs are carried out for commercial clients, e.g. oil companies, engineering consortia etc. While many of these data come into the public domain eventually, they often remain in the grey literature, difficult to access and assess. Yet much of the work is carried out by competent taxonomists and would provide extensive multi-species coverage within certain regions, i.e. North Sea, Alaska, Gulf of Mexico.

Advantages and Disadvantages of Existing Records

ADVANTAGES

1. ***Considerable quantities of data are potentially available.***
   Given that there has been huge collecting effort over the past one hundred years considerable coverage of certain regions such as the N. Atlantic is likely both spatially and temporally.

2. ***Quickly populate models and databases.***
   The main advantage of utilizing existing information is that it can quickly provide data with which to test new models and IT developments.

3. ***Cost-effective data collection.***
   Even data that are not yet available electronically can be rapidly databased. For example at the NHM, as part of the FishBase project, 130,000 specimen records (taxonomic and geographic data) were put on computer in two years, including verification within the collections. The cost of this project was 50K Euros (about $ 50K). One person was employed full-time to carry out the data entry. This is particularly cost effective way of gaining substantial amounts of information in a relatively short space of time.

4. ***Specimens are useful for quality control and ground-truthing distribution models;*** for example, by testing predictions from retrospectively modelling past occurrences and comparing the outcome with known specimen records.

5. ***Existing information represents a pyramid of taxonomic knowledge*** which is now in short supply and will not be replaced quickly or easily.

DISADVANTAGES

1. *Taxonomic consistency*
   There may not be taxonomic consistency across a whole collection within an institute and between collections made to achieve different objectives. Also taxonomic systems change and collections and databases may not be upgraded accordingly.

   The solution would be to make sure there was sufficient taxonomic expertise to provide quality control and assurance through networking and the development of consortia.

2. *New species*
   This is a particular problem for deep-sea and tropical marine biologists, where many new species are encountered. Often due to the time scale of the project, publication of taxonomic results lags

behind. Until such times as the taxonomy catches up, the identification of these species must remain project specific so that it is not possible to reconcile one study's *species A* with another study's *species A*. Several mechanisms have been proposed to try to overcome this problem (Paterson et al. 1999 http://www.nhm.ac.uk/zoology/taxinf). Southern Californian Association of Marine Invertebrate Taxonomists also exchange new taxonomic records using their website (http://www.scamit.org.). Inevitably it will require a group of taxonomists to come together to produce a consistent taxonomy. Many deep-sea data are affected by this problem.

3. ***Data temporally compromised***
Data may cover a wide temporal range. Specimens registered in the NHM range from the 1700s to the present day. It will be important to separate out different time spans when looking at links to environmental parameters.

However, this may also provide valuable temporal records which could be useful in assessing global change. Again this is to some extent a quality issue and by having good taxonomic and environmental expertise it should be possible to provide robust datasets for analysis.

Agreement must be reached on the method by which databases will be linked along with the protocols, terminologies and frameworks which will enable software to extract records from existing datasets.

4. ***Data geographically constrained***
Many early records do not have precise spatial information (Figure 2); specimens often carry only the vaguest geographic information.

There are several ways of coping with these types of data; but ultimately by producing community wide standards, guidance can be given on the utility of specimens lacking geographic references.

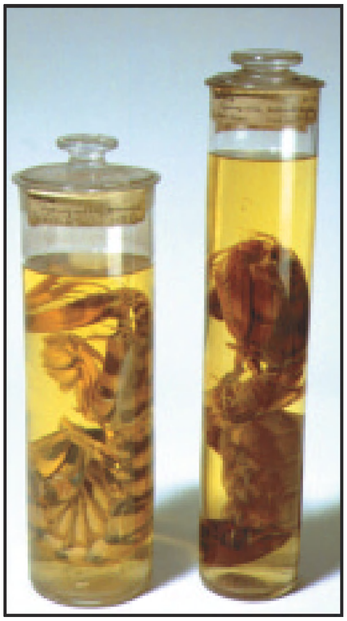

Figure 2. Museum collections have a wide spatial and temporal coverage. These specimens of mantid shrimps were collected in the 18th century. The locality data highlights some of the drawbacks of data derived from historical collections. The one on the left was collected from the Sandwich Islands, while the one on the right was collected from Eastern Seas! (photo courtesy The Natural History Museum)

5. ***Access and intellectual property rights***
These are also important issues when using existing data and, in particular, giving proper acknowledgement to the source of such data.

Many of these issues have been dealt with already and to varying degrees by other data management groups. Physical and chemical oceanographers are experienced in assigning different levels of accuracy and quality to disparate datasets (see references in Proceedings of the International Workshop on Oceanographic biological and Chemical Data Management, 1997). Close liaison between the various communities should resolve many of the potential problems listed above.

Procedural Issues

In seeking to maximise the use of existing data several major issues must be addressed which are procedural rather than technical. Agreement must be reached on the method by which databases will be linked along with the protocols, terminologies and frameworks which will enable software to extract records from existing datasets. These issues of structure, syntax and semantics (the three S's) are being addressed by many groups and it is important that OBIS communicates and coordinates with these groups. For example, Computer Interchange of Museum Information (CIMI: http://www.cimi.org/) has produced a guide to best practice on the use of a series of interchange protocols and standards, including Dublin Core, based on cross-domain agreements, which will make resource discovery more effective. Similarly there are initiatives such as Species 2000, European Register of Marine Species and Integrated Taxonomic Information System (ITIS: http://www.itis.usda.gov/plantproj/itis/index.html) that have faced issues which OBIS will have to solve when the data start to become available. A key issue in producing interoperable databases will be establishing community wide standards. In effect, establishing business rules and classificiations which will provide the mapping between systems, such as synonyms of terms / places / species to be recognized by ETL (Extraction, Transformation and Loading) tools. By gaining agreement on the use of consistent terminology as an aid to precise and comprehensive retrieval a much more flexible approach can be adopted by researchers creating databases of biological information and ultimately savings in both time and resources will result. Organizations such as the International Working Group on Taxonomic Databases (http://www.tdwg.org) are working actively to resolve many of these issues.

Conclusions

The advantages of collating existing knowledge are both scientific and pragmatic. Such data represent the accumulated knowledge of the biology of our oceans as well as an enormous financial investment. They form the foundations upon which we can build the next generation of research.

REFERENCES

Conkright, M.E. and S. Levitus, 1996: Objective analysis of surface chlorophyll data in the northern hemisphere. In: *Proceedings Of The International Workshop On Oceanographic Biological And Chemical Data Management.* NOAA Technical Report NESDIS, 87, 33-43.

Funk, V.A., M.F. Zermoglio and N. Nasir, 1999: Testing the use of specimen collection data and GIS in biodiversity exploration and conservation decision making in Guyana. *Biodiversity & Conservation*, 8, 727-751.

Levitus, S., 1996: Interannual-to-decadal variability of the temperature-salinity structure of the world ocean. In: *Proceedings Of The International Workshop On Oceanographic Biological And Chemical Data Management.* NOAA Technical Report NESDIS, 87, 51-54.

Proceedings Of The International Workshop On Oceanographic Biological And Chemical Data Management. 1997. *NOAA Technical Report NESDIS*, 87, 224pp.

Wüst, G., 1964: The major deep-sea expeditions and research vessels 1873 - 1960. *Progress in Oceanography*, 2, 1-52.

Recommendations

The overall objective for the Census of Marine Life in this area of activity should be to have all known oceanographic specimen-based records databased within 15 years.

This would bring together about 120 years worth of scientific activity.

0-2 years

Objectives:

- Provide Metadata catalog of sources of primary data to identify which institutions hold voucher specimens relating to published cruise reports.
- Gain agreement on community-wide data and terminology standards
- Gain agreement on IPR (Intellectual Property Rights) issues

Actions

Organize a workshop to discuss common databasing issues. Some dialog should be initiated, particularly in the areas relating to data transfer, availability and issues of intellectual property rights, between the Census of Marine Life team and groups such as ICLARM, IODE, other biodiversity researchers such as Dr V. Funk (Smithsonian Institution).

Establish international teams combining taxonomic expertise, databasing professionals and support manpower for databasing, using FishBase as a model. Such teams should be charged to:

> Produce a data catalog listing who has what and in what form. Prioritize key regions known to be rich in data, i.e. N Atlantic (including the Mediterranean), Arctic and Antarctic and determine sources of data.
>
> Prioritize key taxonomic groups, and produce a cohesive database, either dispersed, connected using web technology or as a central system. If necessary provide funds to database particular collections or records to support this effort.

Provide necessary infrastructure to support quality control workshops and establish taxonomy focussed groups to assess data.

Test initial models and information technology (IT) developments.

Create a framework whereby new data being produced could be integrated and made available. This would involve introducing a minimum set of standards which the data generator would have to conform to, but if it was as flexible as possible then this should not be too onerous.

2-5 years

Activities

Prepare submissions to the EU Framework Programme to support European efforts in this initiative.

Expand scope to include wider variety of taxonomic groups in key areas.

Assess existing gaps in knowledge and coverage and prepare research proposals to augment systems.

Carry out quality control and assessment of data

5+years

Activities

Assess status of existing data with potential sources and prioritize to maximize coverage.

Widen number of key areas and again determine sources of data. This could be implemented as a series of discrete projects funded by the U.S. National Oceanographic Program and other agencies.

By 10 + years a substantial amount of existing data should have been databased and available.

Continue collaborative research with modelling and IT developers.