On Preservation of Data

C.P. MATHEWS

Abstract

This paper presents, for three aquatic research projects, the type of data that were collected, and the reasons why these data eventually became lost or inaccessible. A strategy for countering such data loss is proposed.

Introduction

Collection of data is one of the most expensive procedures in both developing and developed countries. This is particularly true for the data required to manage fisheries and aquatic ecosystems where detailed information on growth, mortality, reproductive biology, general biology, catch, effort, landings, prices, supply, demand and other biological and economic data are needed to allow rational management of important stocks, ecosystems and fisheries. During the many years spent largely in tropical countries working on marine and freshwater ecosystems, I have found that inadequate conservation of data over intervals from a few years to several decades has led to avoidable losses of information, and therefore to inefficient use of funds.

Two technical developments have changed the general attitude towards data conservation during the last decade:

a) the likely influence of long-term global changes on ecosystems is now widely understood, and is leading to systematic attempts to conserve and analyze data over long periods of time; and
b) the development of cost-effective methods for electronic storage and processing of large volumes of data is encouraging the creation of large databases, even in developing countries.

In this context the establishment of databases such as FishBase (Froese 1990; Pauly and Froese 1991) is a welcome development in fisheries and aquaculture and will be especially useful in the tropics, although it covers fish of all climatic zones.

But the creation of FishBase and similar databases for taxa other than fish will not by itself be sufficient: I suspect that significant data losses are still the rule rather than the exception, and a change in the attitudes of scientists worldwide is needed. The object of this communication is to show how I addressed (and more frequently failed to address) the question of long-term fisheries data conservation. Some of the problems I confronted were dramatic, but not unique, and workers in tropical countries will sooner or later face some or all of these problems. Hopefully they will do so, armed with ideas as to how they can do better than I did.

England

From 1966 to 1970, I worked at the University of Reading on the International Biological Programme (IBP), on the Dreadnought reach of the River Thames, as part of Ken Mann's IBP team. The main results were reported by Mann et al. (1972) and have been reanalyzed using the ECOPATH II methodology by Mathews (in press) who compared the results obtained from the old IBP approach (Mark 1) with the results obtained using new methods (Mark 2). Although the Mark 1 and 2 models provided essentially similar results, the Mark 2 model was much more complete and exposed a major weakness in the Mark 1 model: the only published biomass estimates are those of Mathews (1971) for fish, and for phytoplankton and fish (Mann et al. 1972). Yet all of the Thames IBP studies provided biomass estimates. Because the IBP work tended to focus attention on production, biomass estimates were obtained but neglected because they were not thought to be useful or interesting. These estimates, had they been conserved, would probably have allowed the identification of a unique Mark 2 model instead of the more tentative one presented by Mathews (in press).

Mathews's (1971) estimate of population fecundity and its contribution to fish production in the Thames was based on a small volume of old data recorded by previous workers and conserved at the University of Reading. Without these data the rather complete study actually carried out would have been impossible. Nevertheless, Mathews's (1971) rather detailed study of the fish populations of the River Thames does not include the raw data used, nor did he take any measures to conserve those data. Nor, when the IBP team was disbanded, was any systematic attempt to conserve the raw data generated during the project.

Mexico

During 1970-73, I worked in Mexico for FAO, and with numerous Mexican counterpart staff I was responsible for surveying about 100,000 nm^2 of sea in the Gulf of California; the resulting research was published (e.g., Mathews et al. 1975; Mathews and Ruiz 1974; Mathews 1975, 1985). The original log books were deposited with the appropriate authority in Mexico City, but so far as I know no further use of the data was ever made.

From 1973 to 1976, I held the Chair of Fisheries Oceanography at a provincial University in Mexico and was in charge of the technical work on fisheries for a major project in the Baja California Peninsula. At that time in Mexico it was customary for serious budgetary problems to accompany a change in government, and this led indirectly to the termination of the contracts of many staff in 1976, and to the collapse of the institutional arrangements in my own department. The immediate result was that the data collected over a period of about 18 months and four major fisheries research expeditions covering about three months' sea time in the study area (Magdalena Bay, Baja California) were never fully analyzed or published. These data were deposited with the appropriate authorities at the University, but appear now to be unavailable.

Kuwait

During 1977-88, I worked at the Kuwait Institute for Scientific Research (KISR). On my arrival, a commercial shrimp trawler converted for use as a research vessel was available, and was used for gathering fish and shrimp broodstock for about six weeks per year. There was no other stock assessment activity at KISR although the culture program led to publications on the biology of some fish. An FAO/KISR project established the foundations for shrimp stock assesment work in 1977-79 (e.g., Mohammed et al. 1981; van Zalinge et al. 1981), and I took over this work from 1979 to 1986, as well as the fin fisheries management work from 1977 to 1988, when I left Kuwait. A total of over 100 reports and 27 papers were published on finfish and about 35 reports and five papers on shrimp and about 35 reports and five papers on shrimp (Mathews 1986a) during the period I was responsible for these projects. These reports and papers were published by the two teams, including scientists from at least 13 countries, with Kuwaiti counterpart staff being the dominant element. A very substantial database was collected and conserved on forms, and was available for use by all scientists. On my departure there was a combined team of about 15 scientists and support staff, and regular advice was being tendered to, and was usually accepted by, the government through annual fisheries management workshops (e.g., Mathews 1986b). A large (450 GT, 42-m stern trawler) custom-built vessel was made available in 1984 as a fisheries and oceanography research vessel and regularly carried out about 130 or more sea days of research per year. An effective catch and effort data collection system was in place and provided fisheries statistics on a regular basis.

The capability developed over the 11 years I was in Kuwait was obviously the product of a magnificent team effort carried out by the KISR management, and the large team of Kuwaiti counterpart staff led successively by Drs. Nizar Husein, Mohammed Al-Attar, Ziad Shehadeh and Mohammed Seif. The core of that capability was the vast amount of data accumulated and tools to analyze those data, based on the small but efficient population dynamics unit developed by Dr. Gary Morgan.

The establishment of electronic databases was started with the preparation of special programs for recording shrimp data (Morgan 1981) and sea survey data (Joseph 1987).

On 2 August 1990, the Iraqi army invaded Kuwait and soon after began to remove equipment, books and data records from Kuwait to the University of Basrah's Marine Science Centre.

Discussion and Conclusions

Causes of data loss

UNAWARENESS OF POTENTIAL USE

A failure to appreciate that today's uninteresting data may be tomorrow's gold mine is a very significant cause of failure to preserve data (Pauly 1992). For instance, data obtained over 20 years ago during the Thames IBP study, e.g., on biomass, were not conserved partly because of lack of interest, and partly because universities are not very suited to the collection and particularly the preservation of large amounts of data: they usually lack the personnel to support mundane data collection and conservation in large amounts.

The same remarks may be applied to the data left behind in the Mexican University, but cannot be true for KISR and the Mexico/FAO Project.

LACK OF FACILITIES

The Mexico/FAO Project work occurred from 1970 to 1973, and only very limited computer facilities were available so that electronic data storage was not feasible. Institutional structures were weak and scientific material (e.g., taxonomic collections, scientific data) were often not preserved because of changes in personnel and priorities which coincided with changes in government. These facts, although foreseeable, were not in fact foreseen, and remedial measures were not taken.

A system for establishment of electronic databases did exist at KISR. For the KISR work it is possible to say in retrospect that I was seriously remiss in not accelerating the establishment of an electronic database.

OWNERSHIP OF DATA

This is perhaps the greatest single cause of withholding, and so of underutilization, of data. A governmental or private organization, having invested considerable sums of money in obtaining information, understandably wishes to ensure that it and no one else profits from that investment. Often, however, this leads to the accumulation, nonuse and subsequent loss of data that could, two to three decades later, become an invaluable baseline or an important source for reanalysis and reinterpretation of a past occurrence/situation. This is a particularly important problem in tropical countries where institutional arrangements may often be weak, and the concern to preserve data for the country's use may be particularly strong.

Note that the oceanographic community withholds for one year analysis of oceanographic data submitted to data centers and that the code of zoological nomenclature similarly allows one year for the discoverer(s) of a new species to describe it.

PERSONAL ATTITUDE TOWARDS DATA

I have noticed in many scientists (not least in myself) a tendency to regard data not just as a source of information but also as a source of influence, power and eventually money. This takes two main forms:

- Conservation of data obtained by a researcher until he or she can publish the analysis is often (understandably) thought to be justifiable, yet has, to my certain knowledge, led to loss/suppression of data, with consequent nonpublication, in many instances and several countries.
- Fear that making available the data upon which a publication is based, is to expose one's self to the dangers of informed criticism through reanalysis, e.g., by professional colleagues or rivals. An uncriticized publication record may be more important for promotion purposes than any impeccably clear set of scientific conclusions achieved after intense public debate.

Remedies

DATABASES

The regional or global storage of information in databases held by international organizations such as FishBase may ensure the survival of data and their conservation in spite of national upheavals, civil wars and other events that are sufficiently common worldwide to warrant consideration by an international data conservation strategy. The widespread use of such databases will ensure that data will be conserved and made available to scientists over intervals of several decades. This would include the dispatch of copies of data to international data conservation organizations outside the country/region of origin so as to conserve data in the event of wars or the failure to ensure institutional continuity and preservation of data for mundane reasons (e.g., budgeting, administrative reform). In the case of FishBase, this need will be met since the database is available to the international community.

PUBLICATION OF DATA

Janzen (1986), Pauly (1987) and Wiechart (1990) have commented on the essential need to conserve and reinterpret old data, and Janzen (1986) suggests that: "There is absolutely no reason why research papers should not also contain a second nonpresentation - Archival Function - as an appendix, as a Table (they are identical) or simply as another section. Set in small type, the cost will be minimal. For editors concerned that their journals be readable in the sense of weekly news magazines, any cursory reader will recognize an archival section for what it is." This approach will ensure that where data cannot conveniently be preserved in a database, it is still preserved for future use.

Pauly (1992) elaborated on this and proposed the concept of "data-rich books". Also in his role as Editor of *Fishbyte*, he has set an example as refreshing as it is rare in requiring that most publications include the length-frequency data upon which ELEFAN assessments are based, so allowing for replication, and reanalysis of the data when new analytical techniques become available (e.g., the application of the GOTCH.A program for reestimating Z for populations from areas with high seasonal temperature variation (Gayanilo 1992)). More actions of this type are needed.

EFFORTS BY INDIVIDUAL SCIENTISTS

It is essential that individual scientists make every effort to ensure that their data are included in publications, or at least in reports deposited in libraries of international institutions. Project Leaders and contracting agencies should insist on this, but action must be started by individuals. If we do not do this ourselves, no one will do it for us.

For individuals working as expatriates, such action may be exceptionally difficult: the need to ensure conservation of data may conflict with the need to ensure confidentiality of data. I personally have always left the data in the hands of appropriate authorities in the countries in which I worked, yet had I been sufficiently dishonest to "steal" copies of the data I would, in retrospect, have done the world's scientific community a small but important service. There is no easy answer to this dilemma: perhaps all countries should be encouraged to share data with a provision for a few years of nonuse so that the scientists who collected them can have the opportunity to analyze and publish them first.

Perhaps this is a matter to be debated publicly: I would welcome a discussion of these matters. Clear guidelines are needed so that marine scientists can take the initiative as individuals acting in different countries and institutions, and so that such initiatives are welcomed by appropriate authorities.

References

Froese, R. 1990. FishBase: an information system to support fisheries and aquaculture research. Fishbyte 8(3):21-24.

Gayanilo, F.C. 1992. A program for constructing length-converted growth curves when growth is seasonal. Fishbyte 9(2):72.

Janzen, D. 1986. Science is forever. Oikos 46:281-283.

Joseph, P.S. 1987. Documentation of Fortran programme for processing sea survey data on fish. KISR Technical Report, No. 2304, Kuwait Institute for Scientific Research, Kuwait.

Mann, K.H., R.H. Britton, A. Kowalczewsky, T.J. Lack, C.P. Mathews and I. MacDonald. 1972. Productivity and energy flow at all trophic levels in the River Thames, England, p. 519-590. *In* Z. Kajoh and A. Hilbricht-Ilkowska (eds.) Proceedings of the IBP-UNESCO Symposium on Productivity Problems of Freshwaters, 6-12 May 1970, Kajimirz-Dolny, Poland.

Mathews, C.P. Productivity and energy flows at all trophic levels in the River Thames, England: Mark 2. *In* V. Christensen and D. Pauly (eds.) Trophic models of aquatic ecosystems. ICLARM Conf. Proc. 26. (In press)

Mathews, C.P. 1971. Contributions of young fish to the River Thames near Reading. J. Fish Biol. 3:157-180.

Mathews, C.P. 1975. Some observations on the ecology and the population dynamics of *Merluccius angustimanus* in the South Gulf of California. J. Fish Biol. 7:83-94.

Mathews, C.P. 1985. Meristic studies of the Gulf of California species of *Merluccius*, with description of a new species. J. Nat. Hist. 19:697-718.

Mathews, C.P. 1986a. Final Report, Shrimp Fisheries Management Project. Kuwait Institute for Scientific Research, Kuwait. 114 p.

Mathews, C.P. 1986b. Conference Report, Shrimp and Fin Fisheries Management Workshop, Kuwait, 11-13 January 1986. Mar. Policy 231-232.

Mathews, C.P. and M.F. Ruiz. 1974. *Cephalurus cephalus*, a small shark taken in the Gulf of California, with a description. Copeia 2:536-550.

Mathews, C.P., J.L. Granados and J. Arvizu. 1975. Results of the exploratory fishing cruises of the Alejandro de Humboldt in the Gulf of California, November 1972. California Commission for Oceanic and Fisheries Investigations Symposium on Waters around Baja California, Yosemite, California.

Mohammed, K.H., M. El-Musa, M. Al-Husseini and A.R. Abdul-Ghafar. 1981. Observations on the biology of the exploited species of shrimp in Kuwait. Kuwait Institute for Scientific Research/Food and Agriculture Organization, Kuwait. 55 p.

Morgan, G.R. 1981. Programme package for shrimp data storage and summarization. Kuwait Institute for Scientific Research, Kuwait. 27 p.

Pauly, D. 1987. On using other people's data. Naga, ICLARM Q. 11(1):6-7.

Pauly, D. 1992. Data-rich books. BioScience 43(3):167-168.

Pauly, D. and R. Froese. 1991. FishBase: assembling data on fish. Naga, ICLARM Q. 14(4):10-11.

van Zalinge, N.P., R. Jones, M. El-Musa, M. Al-Hosseini and A.R. Abdul-Ghafar. 1981. Mark-recapture experiments on the gulf shrimp *Penaeus semisulcatus* De Haan in Kuwait waters. Kuwait Institute for Scientific Research/Food and Agriculture Organization, Kuwait. 66 p.

Weichart, G. 1990. The quality of "old" oceanographic phosphate data. International Council for the Exploration of the Sea, Hydrography Committee C.M. 1990/C:7.

C.P. MATHEWS *is with the Zoology Department, University of Reading, Whiteknights, Reading, England.*