A Strategy and a Structure for a Database on Aquatic Biodiversity*

Rainer Froese
Daniel Pauly

International Center for Living Aquatic Resources Management
MC P.O. Box 2631
0718 Makati, Metro Manila
Philippines

Abstract

In order to improve our understanding of aquatic biodiversity, it is suggested to assemble, in a single database the huge amount of existing data on the occurrence of aquatic species in space and time. Such data are available in museum collections, research vessel surveys, tagging studies, the scientific literature, and a variety of other sources, often in digitized form. The database would be distributed on CD-ROM with annual upgrades. It would preserve data which might otherwise be lost; it would provide baseline data on biodiversity from historic data sets; in combination with data derived from existing biological, oceanographic, and meteorological databases it would allow for analyses of biodiversity which are currently not possible; and it would guide the ongoing efforts towards collection of data that are most useful for analytical models. We suggest to establish a network of institutions that hold relevant data and are willing to share them.

1 Introduction

The need to maintain or restore biodiversity as a precondition to a sustainable use of living resources is now widely accepted. To rapidly improve our understanding of aquatic biodiversity, on a cost-effective basis, we suggest to assemble, in a single database the huge amount of existing data on the occurrence of aquatic species in space and time. Such a database, available on CD-ROM with annual updates would serve a variety of important tasks. Firstly, it would preserve data which otherwise will be lost. Secondly, by combining data from many different sources it will provide an unprecedented coverage over space and time; in combination with existing biological, oceanographic, and meteorological databases it will allow for comprehensive analyses, currently not possible. Thirdly, it is expected that the availability of the database and its analytical tools will guide the ongoing efforts towards collection of those new data that are most useful for analytical models. Thus, the challenge is to establish a network of all institutions that hold relevant data and are willing to share them, to have the data cleaned up (see below) and translated into a common format, and to adapt or develop appropriate models to analyze the data.

---

* ICLARM Contribution No. 1020. Presented at the 6th CODATA/DSAO Meeting, 10-12 March 1994, Taipei, Taiwan.

Once established the database will provide baseline data on biodiversity from the historic data sets. It will allow to identify spatial and temporal trends in biodiversity in the context of, e.g., heavy fishing pressure, pollution, environmental degradation, and global warming.

2 The data sources

Museum collections

Collections of preserved aquatic organisms are maintained in Natural History Museums all over the world totaling probably several million samples and covering a time span from the 18th Century to the present. Each sample usually consists of one or several specimens preserved in alcohol or formalin, and described by a slip of paper stating the species name and where, when, and by whom the specimens have been collected. Several hundred thousand samples are already computerized but the bulk of the data, especially in the many small museums in developing countries, still awaits digitization. The proposed database will be provided for free to these museums to motivate them to digitize their own collections, and to provide a framework for comparisons of their collection with others.

Museum samples, especially the old samples, often stem from expeditions, and detailed reports are normally available with additional information on locality, environmental conditions, species abundance, species caught but not collected, etc. These reports are another important source of information which to date is largely neglected.

There are mainly three problems with museum collection data: firstly, the scientific names are often outdated. In most cases this can be easily repaired with the help of synonymies. Secondly, the locality is often given very vaguely, e.g. as 'Indian Ocean' or as a location in a certain country. The latter can often be translated into coordinates with the help of gazetteers, i.e., books or databases that contain the names and the coordinates of all rivers, lakes, mountains, and places in a country. In both cases it is usually helpful to consult the expedition report or the letters exchanged with the donor of the sample. Thirdly, the date of collection for older samples is often missing. Again the museum files may contain that information. In any case, unclear records can be identified as such and excluded from analysis until their status is clarified.

Research vessel surveys

Research vessel surveys of aquatic organisms have been carried out for more then 100 years with several hundred thousand stations in all oceans. The typical data collected on such surveys are: station data with information on locality, weather, and environment; haul data with number, size range, percent of catch for each species, and length frequencies for selected species; and catch per unit effort. The results are typically presented in aggregated form (e.g. as average over a number of species or stations) in technical reports. Subsets of research surveys are sometimes published as books (Gloerfelt-Tarp and Kailola 1984) or in the primary literature. The raw data sets are rarely published and usually lost after some years (Mathews

1993). The Aquatic Biodiversity Database would try to find and incorporate as many historic datasets as possible; it would offer a convenient archive for new surveys. Raw data of of recent surveys are regularly computerized and some of these files are publicly available (Strømme 1992; Allen and Smith 1988).

The main problem with research vessel data are outdated scientific names and misidentifications. The former can most often be solved by the use of synonymies, the latter by checking the likelihood of occurrence against the established distributional ranges and by consulting the relevant taxonomic literature or databases such as FishBase which do record known misidentifications.

Tagging data

Tagging and releasing aquatic animals is a standard method for estimating the stock size from the ratio between tagged and untagged species that are recaptured. Additional information that can be derived from tagging studies are growth, migratory behavior, and swimming speed. Hundreds of thousands of release and recapture records are available for more than 1,000 species (e.g. Beaumariage and Wittich 1966; Casey et al. 1990; Randall 1962; Stanley 1988). Data for release and recapture typically consist of locality, date, age or length, and weight of individual specimens. Ongoing tagging programs maintain databases. Older studies are available in technical reports only, some of which are already difficult to obtain and which also often contain aggregated data only. Ongoing tagging programs normally use modern databases for storage and analysis. Again the Aquatic Biodiversity Database could act as an archive for these databases which still might be lost once a tagging program ends.

Game fish records

The International Game Fish Association (IGFA) as well as many national associations keep track of the largest (= heaviest) fish caught globally or in their area (IGFA 1991). These 'angling records' are recognized by line class (thickness of the line used) as well as over all tackles. Applicants for angling records have to provide a photo as well as other data such as date, locality, length, weight, and girth height. They have to name independent witnesses confirming that the information provided is correct. The associations verify the identification from the photo, involving experts if necessary. The accepted angling records are published regularly and some are maintained in databases. The main problem with angling records is to translate the localities into coordinates. Again gazetteers can be used and the associations normally know the places and can help. In the biodiversity database such records would contain the name of the contributing association and would be classified as 'based on angling record'.

Scientific literature

The scientific literature is a rich source of biodiversity data, often in the form of revisions, i.e., summarizing and updating all available information for a species group or a region (e.g. Klausewitz and Nielsen 1965; Bruton and Coutouvidis 1991; Pethiyagoda 1991). Since digitizing such information is labor-intensive, we suggest to start with those publications which, according to a set of preset criteria, would contribute most to the goals of the biodiversity database.

Visual census data

Visual census methods (Stoddart and Johannes 1978) are now widely used to estimate the abundance of coral reef fishes (e.g. Russ 1989). Such surveys are usually conducted by well-trained marine biologists and restricted to species that are easy to identify by SCUBA divers because of their unique color patterns. Data are often available in database format. However, because no evidence for correct identification can be provided, such reports have to be matched against a list of species known to occur in the area before records are accepted for the biodiversity database and they would be classified as 'based on visual census'.

Underwater photos from divers

Underwater photos of aquatic organisms taken by divers probably total several hundred thousands, dive magazines giving an impression of the most beautiful photos only. Divers keep log books in which they carefully record the diving site, date, time, maximum depth, and other dive details. Many divers do know the organisms well enough to at least call them by their correct common name. We suggest to invite divers to contribute to the study of aquatic biodiversity by providing underwater photos of aquatic organisms together with the following information: Country, diving site, coordinates, depth, date, common or scientific name, estimated size, estimated abundance, reference used for identification. The procedure for accepting such a record for the database would consist of i) verifying the identification from the photo, involving experts if necessary, and ii) verifying that the locality is within the known distributional range of the species, again involving experts if the record would constitute a range extension. Each record would carry the name of the contributor and would be classified as 'based on underwater photo' thus enabling users to exclude these records from analysis if they prefer to work, e.g., with museum records only.

Existing biological, oceanographic, and meteorological databases

There are quite a number of global biological databases on aquatic organisms in existence or under development. For example, ICLARM holds a large database on fish (FishBase); FAO's Species Identification and Data Programme holds a database on fish, decapods, and cephalopods (SPECIESDAB); the Australian Institute of Marine Science (AIMS) is developing a global database on corals (CoralBase); the World Conservation Monitoring Center (WCMC) holds a database on marine turtles; the Expert Center for Taxonomic Identification (ETI) is developing databases on pelagic mollusc, marine planarians, sponges, and protists. These databases all contain valid scientific names and the established geographic distribution for each species. Some of them contain additional information on morphology, biology, environmental tolerance, etc. Palaeontological studies make use of the oxygen-isotopes ratio in, e.g., coral skeletons to reproduce the coastal water temperature on a seasonal basis for the past centuries (Pätzold 1986). Many oceanographic and meteorological databases are available in the public domain, mostly on CD-ROM. We suggest to establish links between these databases and the biodiversity database.

3 The database

Structure

As opposed to the existing highly elaborated herbarium databases (Pankhurst, 1991), most curatorial databases used for aquatic organisms still consist of one table only (Hureau, 1991). The most elaborated zoological database design that the authors are aware of is MUSE, a curatorial database system for fishes used by several museums in North America and elsewhere. MUSE makes use of relational database techniques to avoid duplication of entries. Thus the information is distributed in four tables, a Primary table which holds information on the specimen with usually one record per sample, a Locality table which contains station data with one record per station/haul only, a Taxonomic History table which keeps track of who has worked on the specimen with potentially several records per specimen, and a Taxonomic Dictionary which validates the generic name and assigns the correct higher taxa to the specimen (Shao et al. 1992). This structure facilitates data entry since station related information has to be entered only once for all the specimens that have been collected there. It also prevents errors in the spelling of the generic name and the assignment to higher taxa, since this information is derived from the Taxonomic Dictionary, a table extracted from Eschmeyer's (1990) database on recent genera of fishes.

It is suggested here to take the relational database design further to accommodate the variety of sources mentioned above, to standardize locality related information, and to accommodate measurements, counts, and other properties derived from the specimens.

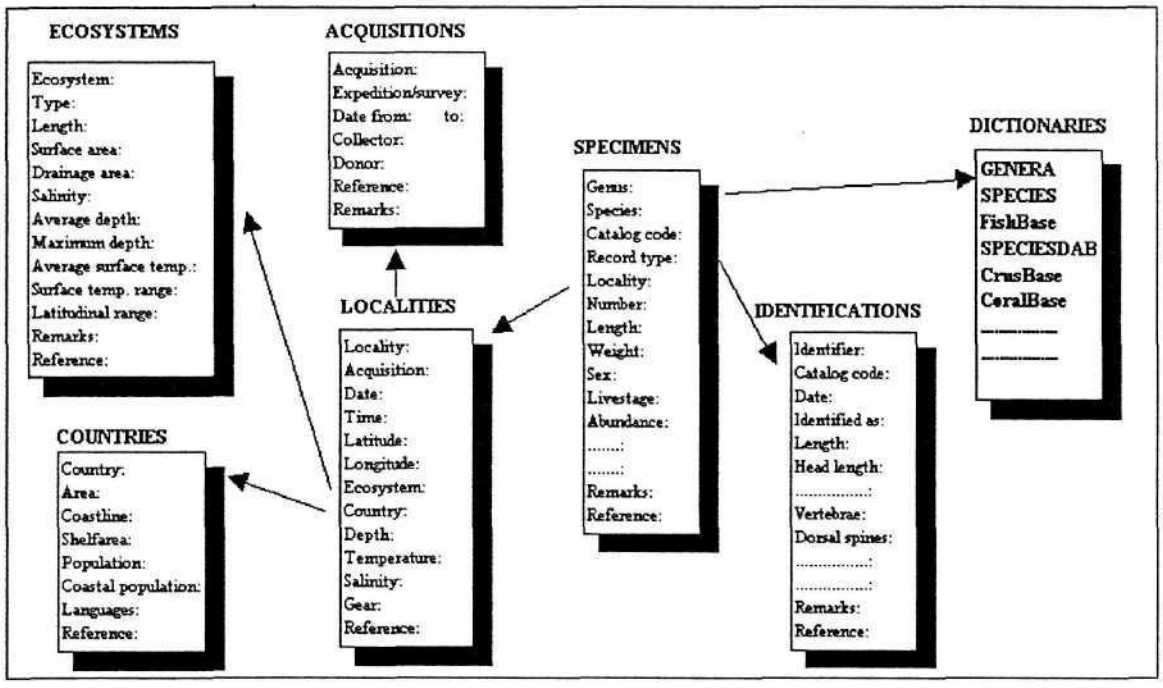

Figure 1 Design of the aquatic biodiversity database. Note that only the main tables, fields, and relationships are displayed.

The suggested design results in at least 6 interlinked tables (Fig. 1), most of which will, however, be hidden from the user. Entries such as scientific name, country, or ecosystem will

be selected from a choice-list thus speeding up data entry and avoiding spelling and typing errors. This concept is also applied to a variety of other descriptive entries which can be easily classified into a limited number of unique choices such as sex, live stage, or type of length measurement. For fish the Taxonomic Dictionary would be Eschmeyer's SPECIES database which is planned to be released in early 1995 and which will allow to validate the specific name in addition to the generic name, thus eliminating another source of errors. For other species groups the above mentioned biological databases can be used as dictionary.

The main reason for the design suggested in Fig. 1, however, is to have better and more reliable access to the data. The structure, for example allows to produce complete lists of species by collector, donor, identifier, station, acquisition, ecosystem, country, and any combination thereof. These lists can be made complete because all of the mentioned selection criteria are selected from choice-lists and thus standardized. Also, if, for example, the name of a country has changed, this is corrected only once in the COUNTRIES table and the change is reflected immediately in all LOCALITIES records.

Table 1 Records found in a North American curatorial database with more than 10,000 records before and after standardization of country names.

Country	before	after	% missing
Bahamas	2	4	50
China	20	67	70
Cuba	1	8	88
Hawaii	20	104	81
Maldives	72	89	19
Myanmar	0	16	100
Philippines	308	330	7
Puerto Rico	143	185	23
Saint Lucia	0	6	100
Taiwan	9	27	67
Thailand	788	1,020	23
USA	337	904	37
Total	1,700	2,670	36

Other convenient outputs from the same data set are: all the surveys and expeditions that have been made in a country, an ecosystem, or a FAO statistical area together with the relevant references; all surveys or expeditions made by a collector; or all the ecosystems in which a taxon has been found. While such lists can also be produced from the existing flat-file tables mentioned above, a search will often fail to find all relevant records because of the lack of standardization. Table 1 shows, as an example, the hits found in a typical flat-file database of more than 10,000 records from a North American museum, before and after standardization of the country names. Overall one third (7-100%) of the relevant records were missed due to name changes (i.e. Burma to Myanmar), different spellings (e.g. Saint Lucia <> St. Lucia, USA <> U.S.A.), absence of standards (e.g. Luzon instead of Philippines), typing errors, and other reasons.

4 The models

A species name, a date, and a locality do not seem to be much of a base for sophisticated analytical models. However, one has to realize that these three bits of data each

represent a vast amount of information: the species name actually provides us with all the biological information on an organism; the locality leads to all the ecological information available for a site; and the date provides us with information about seasonal and historical environmental conditions. In database terms, the species name is the link to biological databases such as FishBase (Froese et al. 1992), the locality is the link to gazetteers, ecosystem databases such as ReefBase (Froese 1992) and to the many oceanographic databases, and the date in combination with the locality is the link to meteorological databases. Since these databases are available on CD-ROM the minimum hardware requirements for sophisticated models that draw on information from all of these sources are a standard PC with a CD-ROM drive. An intelligent software would prompt the user to insert the pertinent CD-ROM, copy all the data needed from this source to the local harddisk, ask for the next CD-ROM, and so on, until all the needed information is assembled and analysis can start. A more advanced system could have an automatic CD-ROM exchanger or multiple CD-ROM drives.

Geographic Information Systems

Biodiversity has to be understood in space and time. The appropriate tool for analyzing spatial data are geographic information systems (GIS). Since all the data in the aquatic biodiversity database will be geo-referenced it can be used for display and analysis by GIS. The biodiversity database will contain MAPPER, a low-level GIS software developed at ICLARM and capable of displaying on screen, e.g., the global distribution of a family or the diversity of certain taxa in different ecosystems (Coronado and Froese 1993). For more sophisticated analyses the biodiversity database can be accessed by commercial GIS such as ARC/INFO or SPANS.

Monitoring ecosystems

The biodiversity database can, for example, be used to monitor the occurrence and non-occurrence of sensitive species, i.e., species that are known to disappear first after a disturbance. For sites where the available samples represent the natural species composition, additional criteria such as percent of omnivores, percent of top-predators, and percent of species depending on high quality habitats for reproduction can be used to derive a general index of biotic integrity similar to the one used for environmental assessment of rivers (Karr, 1981; Oberdorf and Hughes, 1992).

Changes over time in diversity and species composition can be used to identify 'hot-spots' that need special protection or that can be expected to show strong and early signals, e.g., to effects of global warming.

Monitoring species

The Species Survival Commission of the World Conservation Union (IUCN) has developed threatened species categories (from *Low Risk* to *Extinct)* and a set of criteria to place all taxa into at least one of these categories. The criteria refer to absolute numbers and trends in population size, absolute area and trends in distribution, and probability of extinction as derived from models. Reliable estimates of population numbers and areas are much easier to obtain for terrestrial than for aquatic species and therefore the placement of aquatic taxa will depend to a much larger degree on models that also draw on biological, environmental, and

meteorological knowledge, i.e., the information that will be available in the suggested database system.

Thermal niche

It has been shown that, e.g., fish actively try to stay in a 'thermal niche', i.e. a thermal band within +/- 2°C of their preferred temperature. Outside this range their metabolism works less efficiently (Magnuson et al. 1979). Models have been developed to predict the possible effects of climate change on the distribution of aquatic species (Coutant 1990; Magnuson et al. 1990). The biodiversity database will play two important roles in this context: firstly, it will help estimate the preferred temperature which is presently unknown for most species. Secondly, for species with known preferred temperature it will allow to test the predictions of models against the actual distribution pattern over the last 300 years.

Resilience of ecosystems and species

The availability of time series data on population size and area of occurrence or occupancy will allow to study the response of species, species groups, or ecosystems to external disturbances. The resilience or degree of robustness, i.e., how a species or an ecosystem will respond to a disturbance, whether they will return to the former state once the disturbance ends, and if so, how long this will take, is an important piece of information for conservationists and environmental managers. We suggest to develop an indicator of resilience following an approach that has been suggested by Lightfoot et al. (1993).

5 The strategy

Cleaning-up and assembling in a single database all existing data on aquatic biodiversity is such a huge task that any success within a reasonable time frame largely depends on how the task is approached. We suggest the following strategy:

Create a relatively small core group consisting of a modeller, an aquatic biologist, a network coordinator, 3 research assistants, a programmer, and a secretary. This group forms the center of a network of institutions holding the above mentioned type of data.

The core group provides network members with regular updates of the complete biodiversity database and with tools to analyze the data. In exchange, the network members make their data available for inclusion in the biodiversity database.

The core group provides institutions with not-yet digitized data with the appropriate database software. If necessary, it assists them in applying for funds for hardware and personnel.

Every record in the database will carry the 'stamp' of the contributing institution thus giving visible credit to the supplier of the data.

Institutions that are concerned about others using their data might obtain a lead time of, e.g., 6 month, before their data are made publicly available.

Among the network members several institutions take on the role of coordinators for groups of aquatic organisms such as fish, crustacean, cephalopods, marine mammals, etc. They coordinate the activities of institutions holding data on these taxonomic groups. They assist these institutions in cleaning-up and standardizing their databases and verify the reliability of the data (see below) before they pass it on to the core group for inclusion in the biodiversity database.

The core group produces a CD-ROM containing the biodiversity database as well as analytical tools. Regular updates of the CD-ROM will be distributed for free to network members. Others will be able to purchase the CD-ROM for less than 100 US$.

The core group publishes - in collaboration with the coordinating institutions - a network newsletter to keep members informed of progress, problems and solutions.

The core group maintains a mailbox on an appropriate E-mail network to facilitate communication and data transfer and to provide on-line access to the biodiversity database and the latest tools.

The only efficient and reliable method to verify the content of large databases is to electronically compare them with other, independent databases and to manually verify the mismatches. This will be the task of the coordinating institutions with assistance from the core group. A number of such independent databases have been presented above and the holding institutions might be asked to become coordinators. Coordinating institutions will need at least a scientist and a research assistant to fulfill their task. If necessary the core group will assist them in finding funds for these positions.

We suggest to start this exercise with a project which, if successful, will be turned into a permanent activity of an appropriate international body, similar to the institutionalized gathering of meteorological, oceanographic, and recently also coral reef data.

6 References

Allen, M.J. and G.B. Smith. 1988. Atlas and zoogeography of common fishes in the Bering Sea and Northeastern Pacific. NOAA Technical Report NMFS 66: 151 p.

Beaumariage, D.S. and A.C. Wittich. 1966. Returns from the 1964 Schlitz tagging program. Florida Board of Conservation. Technical Series No. 47: 50 p.

Casey, J., H.W. Pratt, N. Kohler and C. Stillwell. 1990. The shark tagger 1990 summary. Newsletter of the Cooperative Shark Tagging Program. NOAA, Rhode Island. 12 p.

Coronado, G. and R. Froese. 1993. MAPPER, a low-level geographic information system. NAGA 16(4):43-45

Coutant, C.C. 1990. Temperature-oxygen habitat for freshwater and coastal striped bass in a changing climate. Trans. Am. Fish. Soc. 119: 240-253

Eschmeyer, W.N. 1990. Catalog of the genera of recent fishes. California Academy of Sciences, San Francisco. 697 p.

Froese, R. 1992a. REEFBASE - a global database of coral reef systems and their resources. ICLARM Conference Proceedings, in press.

Froese, R., M.L.D. Palomares and D. Pauly. 1992. Draft user's manual of FishBase. Software 7. International Center for Living Aquatic Resources Management, Manila, Philippines. 56 p.

Gloerfelt-Tarp, T. and P.J. Kailola. 1984. Trawled fishes of southern Indonesia and Northwestern Australia. ADAB, Australia, DGF, Indonesia, GTZ, Germany. 406 p.

Hureau, J.C. 1991. La base de données GICIML Gestion informatisée de collections ichthyologiques du Museum, p.225-227. *In* Atlas Preliminaire des Poissons d'Eaux Douce de France. Conseil Supérieur de la Pêche, Ministère de l'Environment, CEMAGREF et le Museum Nationale d'Histoire Naturelle, Paris, 232 p.

IGFA. 1991. World record game fishes. The International Game Fish Association. Fort Lauderdale, Florida.

Karr, J.R. 1980. Assessment of biotic integrity using fish communities. Fisheries 6(6): 21-27

Klausewitz, W. and J.G. Nielsen. 1965. On Forsskål's collection of fishes in the Zoological Museum of Copenhagen. Spolia Zoologica Musei Hauniensis XXII. Copenhagen. 67 p.

Lightfoot, C., M.A. Bimbao. P.T. Dalsgaard, and R.S.V. Pullin. 1993. Aquaculture and sustainability through integrated resource management. Outlook in Aquaculture. 22(3): 143-150

Magnuson. J.J., L.B. Crowder, and P.A. Medvick. 1979. Temperature as an ecological resource. American Zoologist 19: 331-343

Magnuson, J.J., J.D. Meisner and D.K. Hill. 1990. Potential changes in the thermal habitat of Great Lakes fish after global climate warming. Trans. Am. Fish. Soc. 119: 254-264

Mathews, C.P. 1993. On preservation of data. Naga 16(2-3): 39-41

Oberdorff, T. and R.M. Hughes. 1992. Modification of an index of biotic integrity based on fish assemblages to characterize rivers of the Seme Basin, France. Hydrobiologia 228:117-130

Pätzold, J. 1986. Temperatur- und CO_2-Änderungen im tropischen Oberflächenwasser der Philippinen während der letzten 120 Jahre: Speicherung in stabilen Isotopen hermatyper Korallen. Berichte-Reports, Geol.-Paläont. Inst. Univ. Kiel. Nr. 12. 92 p.

Pankhurst, R.F. 1991. Practical taxonomic computing. Cambridge University Press. Cambridge, 202 p.

Pethiyagoda, R. 1991. Freshwater fishes of Sri Lanka. Wildlife Heritage Trust of Sri Lanka. Colombo. 362 p.

Randall, J.E. 1962. Tagging reef fishes on the Virgin Islands. Proc. Gulf and Caribb. Fish. Inst. 14: 201-241

Russ, G.R. 1989. Distribution and abundance of coral reef fishes in the Sulimon Island Reserve, central Philippines, after nine years of protection from fishing. Asian Marine Biology 6: 59-71

Shao, K.T., L.S. Chen and L.Y. Hsieh. 1992. Current status of fish databases of Taiwan. p.83-96 In: J.L. Wu and C.P. Chen (eds.), Proceedings of Workshop on Information Management of Zoo-resources in Taiwan. Academia Sinica, Taipei, Taiwan.

Stanley, C.A. 1988. Tagging experiments on Australian salmon (Arripis trutta): recapture data for Tasmanian releases, 1949 to 1964. CSIRO Marine Laboratories, Hobart, Report 193: 57 p.

Stoddart, D.R. and R.E. Johannes (eds.). 1978. Coral reefs research methods. Monografs on oceanographic methodology. UNESCO, Paris. 581 p.

Strømme, T. 1992. NAN-SIS: software for fishery survey data logging and analysis. User's manual. FAO Computerized Information Series (Fisheries) No. 4, Rome, FAO. 13 p.