SPC/Fisheries 19/WP.2
15 June 1987

ORIGINAL: ENGLISH

SOUTH PACIFIC COMMISSION

NINETEENTH REGIONAL TECHNICAL MEETING ON FISHERIES
(Noumea, New Caledonia, 3 – 7 August 1987)

Towards a Regional Pelagic Fisheries Data Base

(Paper Prepared by the Secretariat)

Introduction

1. The concept of a Regional Pelagic Fisheries Data Base (RPF_{db}) has been discussed frequently and is generally supported by all interested parties. Nevertheless, technical implications have not received a full discussion. In particular, the contents, applications, and mechanism of implementation of such a data base have not been clearly articulated.

2. From the start it must be very clearly emphasized that a data base is not simply a collection of facts. It is much more than a heap of data sitting in a file – whether on a computer or in a desk drawer. A data base includes not only data but also procedures for data acquisition, collation, retrieval and reporting. These aspects of data base development are discussed below.

3. The South Pacific Commission Tuna and Billfish Assessment Programme has been developing one component of a RPF_{db} for several years. The experience gained in this effort is discussed below and should be of value in guiding future development.

Components

4. Components of a data base obviously depend on intended use of the system. Various uses of a RPF_{db} have been discussed in different contexts and include both "management" and "stock evaluation". The concept of management is applied in its broadest sense and includes activities ranging from access negotiation to surveillance. Table 1 indicates some of the possible information components previously mentioned for inclusion in a RPF_{db}. Also included are realistic estimates of the time lag between the event reported and data access by potential users of the information. This time lag reflects the nature of the fishing process and may often restrict the usage of information.

TABLE 1. List of data sources and information components of a Regional Pelagic Fisheries Data Base. The column "Age" indicates the intrinsic time lag between the occurrence of an event and its receipt by the processing centre under normal conditions. The column "Use" indicates general areas where the information in the data base might be applied.

Data Source	*Component*	*Age*	*Use*
Surveillance	Sightings	1 day	Enforcement; Validation
Telex Reports	EEZ Entry	1 day	Enforcement
	EEZ Exit	1 day	Enforcement
	Weekly	1 – 7 days	Enforcement; Conservation
Log Sheets	Catch & Effort	3 – 6 months	Stock Evaluation; Revenue Estimation
Observer Reports	Surveillance	3 months	Enforcement
	Daily Activities		Validation; Effort Estimation
	Biological Data		Stock Evaluation; Revenue Estimation
Port Sampling	Landings	1 week	Revenue Estimation; Validation
	Biological Data	3 months	Stock Evaluation
Research Cruises	Various	6 – 12 months	Stock Evaluation
Regional Register	Various		Enforcement; Licensing; Validation

5. **Surveillance Reports.** These originate from aircraft or surface vessel reports and would normally be available to a central area within a few hours of observation. The principal use of this information would be for enforcement of access agreements and, potentially, fleet management.

6. **Telex Reports.** Vessels are required to report their catch on board by telex weekly and when entering and leaving EEZs. Information could be available within 1 to 7 days. These reports are applicable to enforcement, fleet management and could also be used for data validation and justification (see section on Processes below). Telex reports could also be employed for monitoring catch in relation to quotas imposed for stock conservation or other purposes.

7. **Daily Log Sheets.** Fishing activities (including position, effort, and catch) are recorded daily on standard forms and forwarded to countries and then eventually to the SPC for processing. There is usually a 3- to 6-month lag between the actual fishing day and data acquisition. Delays are caused by length of trip, time spent coding information in country of origin, postal delay to the region, delay in local fisheries offices, postal delay to central processing location, and actual data processing. Applications of log sheet data include stock evaluation and calculation of catch per trip for revenue estimation. Log sheet data are considered by many to be potentially the most useful and reliable source of information about the fishery. Attempts to assess the reliability of daily catch reports have been discouraging in terms of both accuracy and coverage.

8. **Observer Reports.** On-board fisheries observers have the opportunity to record information usually not available on log sheets (eg. searching activities) as well as to verify the data on the daily log sheets. Species composition of the catch and, in some cases, size composition may be verified by actual sampling. On-board observers may also have the opportunity of making surveillance-type observations if their vessel is operating among other fishing vessels. Again the time lag between observation and acquisition depends on trip length and in-country delays.

9. **Port Sampling.** Data from port sampling might include data gleaned from fishing logs, sales slips, and other documents, as well as from biological sampling of catch. Sales slip information could be used for price calculation and for validation of log sheets. Biological sampling at port of unloading is an alternative to on-board sampling by fisheries observers and applications of the information are similar. The time lag between the event recorded and data acquisition may be as short as a few days for sales slip information to as long as several months for biological data.

10. **Vessel Register.** The regional register of fishing vessels is compiled as licences are issued. It is useful for enforcement, fleet management, and data validation (see section below).

11. **Research Cruises.** Various institutions (eg. Distant Water Fishing Nations, regional organisations, national fisheries offices) conduct research programmes related to pelagic fisheries in the region. The results of these programmes are often available in the form of research and data reports from which biological information may be extracted and used for various purposes, such as stock assessment and data validation.

Processes

12. Each component of the RPF_{db} must pass through a set of rigorous processes prior to being fully incorporated. These processes distinguish a data base from a heap of data.

13. **Preliminary Steps.** Prior to actual creation of a data base the contents and uses of the data and the costs of gathering the data must be clearly understood. Information can only be retrieved from a system if it has been included. Potential users of the system must be consulted to ensure that the information they require will be included in an appropriate manner. Data collection costs restrict the amount of data that can be collected and it is sometimes possible to calculate the relative costs of estimating certain factors to a satisfactory level of accuracy. The calculation may indicate which information is worth collecting and which information should be obtained by other means.

14. Design of forms is an important step in data base development because it ensures that the users get what they need and that the information is reported in a consistent manner in a format acceptable to those who actually complete the form. These activities should be reviewed periodically since requirements change with experience and politics.

15. **Document Inventory.** The path between recording an observation and its incorporation into a data base is often tortuous. When data are recorded at sea, whether by fishermen or observers, the path is even more tortuous. Usually the information is recorded on a piece of paper, but in the future, the use of computer-readable media and direct electronic data transmission will increase. Whether it is a piece of paper or a microcomputer diskette, the original item is called the "source document". A meticulous accounting system is required to track source documents from their origin through all subsequent processes. The problems of document control increase geometrically as the volume and diversity of source documents increases. The importance of having reliable document control procedures cannot be overemphasized in the operation of a complete data base management system. The document inventory process requires establishment of inventory systems at all levels of document transmission. In other words, for documents supplied by fisheries officers within the region, inventory processes need to be established in–country as well as in the processing centres. This procedure will assist in data correction and location of lost data and will reduce the possibility of duplication of data.

16. **Encoding.** It is often the case that the information has been recorded on the source document in a format different than stipulated, ie. entering "December" instead of "12" for month. In some cases these discrepancies are easily detected and repaired by human intervention prior to the data entry process.

17. **Data Entry.** After the information on the source document has been correctly coded at the processing site, it is copied *verbatim* into a computer under the control of a data entry program. For a paper document this process involves the labour of a data entry clerk. For other media, some computer input utility is invoked.

18. **Verification.** After the data have been entered into the computer, the same information is entered a second time, and any differences between the first and second entry are noted and corrected. For paper documents, this process involves the labour of a second data entry clerk and two different people should do the entry and verification in order to achieve the highest accuracy. For other media, most input utilities include the option to reread the data file and ensure that the copy is exact. The verification process, whether human or machine, ensures that the information actually in the computer is exactly the same as on the source document. It does not ensure that the information (on the source document) is correct.

19. **Validation.** Many simple errors, such as misplaced decimal points and miscoded date fields may be easily detected by application of simple range-checking procedures, ie. ensuring that skipjack are less than 20 kg in weight or that there are less than 31 days fished per month. Other simple errors can be detected by comparison with outside information, ie. vessel name may be checked against a list of valid names to detect mis-spellings. Validation is often most easily accomplished as a component of the data entry process.

20. **Standardization.** The same information is sometimes reported in different formats when supplied by different countries, under the terms of different agreements, or as data reporting requirements change. Different formats must be made compatible either by rewriting data into the desired format (translation) or by the use of more "intelligent" software that is able to find the desired information wherever it is.

21. **Justification.** Various components of the data base must be consistent with one another. It must be possible to check whether the telex reports of catch on board, daily catch records, observer reports, and port sampling data agree with one another. The justification process puts diverse pieces of information, most of the components listed in Table 1, into a procedure which allows comparisons and corrections to be made and anomalies noted.

22. **Summarization and Reporting.** The individual records of each component of the data base are often of limited value in their basic form. Data records with similar attributes, eg. deriving from the same vessel or from the same EEZ, must be combined into a more meaningful form. Summarized data must be reported in a format of value to the ultimate user of the data base. Often it is desirable to have a set of standard summaries produced at regular intervals in order that fisheries managers have information which is as up-to-date as possible. Depending on the specific component, these summaries may be required at different intervals ranging from daily to yearly. In addition to such standard summaries, a variety of *ad hoc* summaries will be required for specific analyses. Furthermore, subsets of the data will be required for specialized purposes. These subsets may often be extensive and contain modifications that cause them to diverge from the original data. For instance, the raw estimates of fishing effort may be replaced by some effort measure which corrects for vessel size or alternative average weight estimates may be substituted for reported values. The distinction between such data summaries and the original data must always be made clear.

23. **Access and Distribution.** Data are only valuable to the people who need to use it and an important aspect of data base development is providing data users with access to the data base. Again, depending on the timeliness of the component, access may be needed on a daily basis requiring "on–line" access for the user. Alternatively, it may only be necessary to post summarized information through the mail system in the form of both printed documents and computer–readable media. As with the summarization and reporting process, a clear distinction must be made between subsets of the data base in decentralized locations, divorced from updating procedures, and the on-line data base. Only one data base may be considered definitive.

24. **Documentation.** Users of the data base must know exactly what is in it and how it got to be there. Each component must be thoroughly described in such a way that users will know how the data were collected initially. Each process involved must also be thoroughly described so that the modifications to the data components are clear and users can actually use the system. The documentation should in fact be a user's manual for the data base and be a means of introducing new users to the data base.

Current status of Regional Pelagic Fisheries Data Base

25. Components of the RPF_{db} are maintained at both the South Pacific Commission and the Forum Fisheries Agency. The SPC Tuna and Billfish Assessment Programme currently carries out the task of generating and maintaining components of the RPF_{db} derived from daily log sheets. Small amounts of biological information in the form of length frequency distributions are accumulating at the SPC but not in a formal data base framework. Similarly, a growing volume of data from observer programmes (both by national and SPC observers) and from DWFN research cruises is accumulating in TBAP files. Work is in progress at the SPC on the preliminary phases of acquiring and coding data from both observers and port sampling. A copy of the Regional Vessel Register is also available at the SPC. The FFA generates and maintains several other components of the RPF_{db} such as the vessel register and a certain amount of telex information.

26. The current division of the RPF_{db} seems to be along the lines of timeliness of data access and area of usefulness. The Forum Fisheries Agency maintains components to which access may be required on a day–to–day basis for application to "management", whereas the South Pacific Commission maintains components that are only available after a significant time lag for application to "stock evaluation".

27. Table 2 reviews SPC work on the daily log sheet component of the RPF_{db} in relation to the processes discussed above. A similar table could be prepared pertaining to all components of the RPF_{db} including those maintained at the FFA. The symbol ++ indicates that the process has been implemented in a satisfactory manner; + indicates that the implementation has been accomplished, but is not satisfactory; — indicates that the process has not been implemented or is completely unsatisfactory. These judgements are technical evaluations by the statistics staff of the Tuna Programme based on daily contact with the data. Data users may have a different point of view.

28. **Preliminary Steps.** In the initial stages of establishing the log sheet processing mechanism at the SPC there was insufficient attention given to the complete range of potential uses of the data. The result was a set of problems which have persisted. The situation has improved over the past two years and these problems have largely been eliminated. End uses are more clearly defined and data base contents have been adjusted accordingly. Design of forms has been an ongoing activity and the currently used forms provide most of the information required for management, but lack some important stock evaluation information. New forms have been designed which improve the clarity of information reported and refine data collected for stock evaluation. There have apparently been political as well as technical obstacles to the full adoption and use of these new forms.

TABLE 2. Review of the process implementation on daily log sheet (catch and effort) component of the Regional Pelagic Fisheries Data Base at the South Pacific Commission. See text for explanation of symbols.

Process	*Implementation*
Preliminaries	+
Document Inventory	+
Encoding	++
Data Entry	++
Verification	++
Validation	
internal	++
external	+
Standardization	++
Justification	—
Summarization & Reporting	++
Access & Distribution	++
Documentation	+

29. **Document Inventory.** The SPC makes a list of all documents received and returns this list to the country that supplied the data. As each batch of documents is processed, its progress through the process is carefully monitored and recorded. Once documents have been received and acknowledged by the SPC, the internal document inventory process is satisfactory. There have been some attempts to implement in–country document inventory systems, but the general problem has not been approached systematically. Further work is required so that in–country document inventory systems are both useful in–country and consistent with SPC practices. Such a system could be logically coupled to some in–country preprocessing where such a requirement exists.

30. **Encoding.** Encoding is a routine task normally done by a data entry clerk with several years of experience in handling log sheets. Anomalies are recognized and brought to the attention of supervisory staff, normally a data base specialist or biologist.

31. **Entry and Verification.** These two processes are done under the control of computer programs developed by TBAP staff. These programs control the screen of the computer terminal and bring errors to the attention of the data entry operator. Normally there are two clerks involved, one for entry and one for verification, who alternate tasks. The data entry and validation process depends on the format of the source document and must be modified for each document type received. Source documents in over 30 different formats have been processed at the SPC and about 60,000 daily fishing records are processed annually. Although the data entry and verification processes are satisfactory, as judged by the volume of data processed and fidelity to the contents of the documents, there are serious flaws in the present software. From the restricted standpoint of volume and accuracy, the data entry and verification process is satisfactory. There is, however, no satisfactory update procedure and some validation procedures require separate steps.

32. **Validation.** At the SPC, validation is a two-phase process. *Internal* validation (ie. rudimentary range checking) is built into the data entry process and is satisfactory. *External* validation (eg. checking vessel spelling) is done by separate software. External validation is generally difficult to implement because the range of potential errors is not known until most of them are made. Furthermore, it is a task that taxes both hardware and software and, until recently, the SPC lacked the tools to do the job properly. Lack of complete data validation procedures has caused errors in summary reports that are corrected on an *ad hoc* basis.

33. **Standardization.** All data must be standardized into a common format prior to summarization and reporting. Standardization is carried out by a specialized computer program for each source document format. The standardized daily catch records currently occupy about 95 megabytes of mass storage and increase in volume at the rate of 5 – 7 megabytes per year.

34. **Justification.** Preliminary attempts have been made at the South Pacific Commission to justify daily catch records with other components of the RPF_{db}. The regional register is used to assist in validation of vessel names. Certain *ad hoc* comparisons have been made to FAO statistics, observer reports, and other published sources for estimating coverage. The results have been discouraging, both from the standpoint of the difficulty of the job and the lack of correspondence between data sources. Implementation of a formal justification procedure remains a major challenge to those involved in creating the RPF_{db}.

35. **Summarization and Reporting.** Catch and effort data from daily logsheets are routinely summarized by country, by trip and by port of landing. These summaries are produced quarterly and forwarded to countries and to the Forum Fisheries Agency as paper print-outs and, in some cases, in machine–readable media (magnetic tape and microcomputer diskette). In addition, a data base, summarized by 1-degree geographic square, is maintained. This data base is keyed by gear type, year, longitude, latitude, month, vessel size class, vessel nationality, and country supplying data. Information published by Japanese, Korean and Taiwanese national fisheries administrations is also maintained in the same format so that long time series may be examined. The geographic catch and effort data base currently occupies about 45 megabytes of disk volume.

36. **Access and Distribution.** Normal data access and distribution is accomplished by the quarterly summary process mentioned above. For special analyses, access to Tuna and Billfish Assessment Programme data is usually by written request followed by mailed printed reports. Data are considered confidential and access to daily catch records is only given with the prior consent of the agency that forwarded the data to the SPC initially. Summarized data is often used in general publications in the form of maps of catch distribution and plots of trends.

37. On occasions where a deadline must be met, data are transmitted over commercial telephone links between computers in Noumea and Honiara. The method of data transfer is routine in most areas of the world, but was found to be difficult and expensive between Honiara and Noumea. The cause of the difficulties was due entirely to the sub-standard communications networks in place. Up–to–date equipment has recently been installed in Noumea by P.T.T., and should similar equipment become available in Honiara, direct computer–to–computer linkages could become routine.

38. **Documentation.** The Tuna Programme maintains a document which describes steps involved in processing daily log sheets and related information. The format of each computer file is tabulated in detail. Various retrieval programmes and their use are also described. This document is revised periodically and serves as a bridge in staff turnover. The principal problem with this document has been that its priority has been low relative to other activities and is not always completely up to date.

Difficulties

39. Development of the log sheet component of the RPF_{db} at the SPC has not been without difficulties. These difficulties may be traced to a small number of root causes. A clear understanding of these causes may assist in avoiding similar problems in future development of the RPF_{db}.

- Misunderstanding of Purpose. During the initial phases of data base development at the SPC, the purpose, and more importantly, the uses of the data base were not clearly stated.

- Underestimation of Work. All concerned with development of the data base at the SPC have underestimated the work required. Initially there was one data entry clerk and one fisheries scientist, and effort was concentrated on entry and reporting processes. The result was more a heap of data than a data base. The situation has improved markedly in the last 2 years, but the workload still taxes a staff of about $4\frac{1}{2}$ (2 data entry clerks, one data base specialist, one fisheries scientist, and a part-time programmer).

- Misalignment of Staff. In designing a data base, a data base specialist with some fisheries experience is as important as a fisheries biologist who is adept with computers.

- Lack of Proper Tools. The HP1000 computer is a poor choice for data base development. It is not sufficiently powerful (lacking memory), the operating system is hostile to system development, and high quality data base management software is not available.

It should be a simple matter in the future to avoid repetition of these problems.

The Spectre of Duplication

40. The possibility that the Forum Fisheries Agency may attempt to duplicate some of the data base functions currently carried out at the South Pacific Commission has been raised on several occasions. Such a duplication would have consequences beyond the misuse of resources. An attempt to create a duplicate data base from source documents would be technically foolish and would ultimately reduce the credibility attached to both the SPC and the FFA. There are numerous problems of judgement and interpretation required to create a data base, such as interpretation of proper names of ports and vessels and handling of missing dates or weights on log sheets. It would be impossible to ensure that the contents of the two data bases would be identical.

41. Discrepancies have already occurred in respect to missing weight information. The general South Pacific Commission policy is not to introduce external estimates of average weight into reports unless specifically requested to do so. Such average weight estimates have nevertheless been introduced by other agencies into Tuna Programme reports. As a consequence, there are three separate — and quite different — estimates of catch per trip for certain subsets of the data. The individuals who introduced the external average weight estimates have left the region and it is difficult to verify the authenticity of the estimates. The consequence of data base duplication is the loss of the most important reason to have a data base in the first place — credibility. It is essential to be able to state with confidence that the data base accurately reflects the events in the fishery as reported.

42. Users of the data will of course generate analyses and reports for specific purposes. That is what a data base is for. But there can be only one definitive, authoritative data base.

43. Participants in the data base often have legitimate reasons to insist on local implementation of certain processes or to maintain certain data components. These reasons usually hinge on questions of speed of access. Such requirements pose problems to which there are technical solutions that do not require duplication of the data base or even duplication of effort.

Next Steps

44. The development of a RPF_{db} is critical for orderly expansion of pelagic fisheries in the near future and sustained exploitation of these fisheries over the long term. The unique position of tuna fisheries in many Pacific Island states confers special importance to development of a useful data base of the highest quality.

45. The steps necessary to avoid the problems articulated above are clear.

1. In the years since the inception of data base development activities at both SPC and FFA, the intended uses of the data have become clarified. These uses should be formally reviewed and enumerated.

2. Tables 1 and 2 need to be completed and expanded. Both the list of data components and the list of processes need to be examined in relation to projected uses to ensure they are complete. Then the status of the existing implementation of each process on each component should be reviewed. These steps will clearly identify progress to date and work to be done in the future. Work can be divided into component tasks with assigned priorities and deadlines.

3. The utility of a data base depends on the existence of suitable hardware and software and on the ability to access the system. The existing hardware and software at the SPC, the FFA and in member countries needs to be reviewed. Standards for selection of hardware and software should be set.

4. On the basis of this analysis, an appropriate work schedule needs to be developed which clearly identifies the responsibilities of all participating parties — South Pacific Commission, Forum Fisheries Agency, and member countries. To date, data base development has proceeded in each of these three areas with little coordination of activities. In the future, activities must be closely coordinated to ensure that the data base actually does what its users expect.

5. Steps need to be taken to upgrade telecommunication channels between sites and between countries; a first step is installation of locally accessible packet switching ports such as are commonly available throughout most of the world (eg. TOMPAC and GTE/Telenet).

46. The development of a RPF_{db} is a complex and expensive task. We can ill afford misdirection and duplication of effort. The current division of activities between the South Pacific Commission and the Forum Fisheries Agency is based on logical subdivision of the information and can be turned to productive advantage provided there is coordination and cooperation. Division of labour does not require duplication of effort.