

Reportonthe Joint Session to consider Data Handling Problems

1. The Joint Session, convened on Thursday, 9 October, was co-chaired by the Chairmen of the Statistics and of the Hydrography Committees.
2. Contributions presented could be largely grouped into two main categories:
 - 1) those dealing with the establishment of data base management systems, and
 - 2) those concerned with techniques for coping with large data sets.
3. The Joint Session discussed various approaches to cope with the data explosion in marine science in member countries by means of intelligent data gathering, automated data handling and interdisciplinary data structuring. Documents C.M.1980/D:12 and C.M.1980/D:13 underlined, inter alia, the necessity under the present conditions of rapidly increasing data volumes collected under various projects, of selective data acquisition for storing and further processing and of the proper maintenance of data banks. In reply to the questions, a description of the North-East Monitoring Program (NEMP) was provided the aim of which is to provide a systematic time series of observations of components of the marine ecosystem, so as to determine existing levels, trends and natural variations in these components, so that unnatural, and possibly destructive, variations may be determined. Data sets under the program are to be collected from a variety of disciplines, including physiology, pathobiology, genetics, benthos, chemistry, oceanography and fisheries. The Session recognized the high value for potential users of the description of a system of onboard laboratory automation developed for the Danish research vessel "Dana" given in Document C.M.1980/D:20. The system provides a flexible concentration mechanism for the multi-disciplin data acquisition and registration. It is composed of three interconnected processors each equipped with a special-purpose set of peripheral units. Front-end systems to the central computer are comprised of a laboratory (or "chemistry") computer providing for on-line data acquisition from transducers and laboratory equipment and of an acoustic computer acquiring data on-line from the scientific echo-sounder equipment. Data received from the front-end processors and terminals of the central computer are stored in "stations" of a number of disc-resident databases which, when filled, are transferred to magnetic tapes to provide disc space for the creation of new databases. The Session was interested to learn of an experimental fisheries-oriented database using the PASCAL/R programming language/database management system, which was recently tried at the BIOMASS data workshop in Hamburg (Doc.C.M.1980/D:4). The Session noted that such high-level multi-disciplinary systems as PASCAL/R, have an information, rather than data, retrieval function, and that they have reached a level of development which enables them to be used at workshops to provide interactive co-operative interpretation of data brought to the

workshop by participants. This is particularly valuable with regard to such types of data which are not usually available, or not exchanged, through existing data exchange systems. The relative cost of the BIOMASS-type workshop is estimated to be within 0.04-0.08% of the cost of acquiring data. The Session agreed that the prevailing trend in the near future would be from data base management to "information" management systems, which should facilitate for the user the retrieval of the specific information required.

4. Under the second category were discussed some statistical methods for studying and analyzing extensive sets of data, both such methods suitable for calculation by hand and such requiring access to a computer (Doc.C.M.1980/D:10). The author warned against the uncritical amassing of observations just because they are easy to make and store (on computers). He emphasized the need for looking at the data in a meaningful way and not just rush to the computer. It is important not to lose touch with the data. A lot of useful information could be obtained from the data by primitive means (plotting, smoothing techniques).

Techniques for reducing large volumes of oceanographic data were dealt with in Doc.C.M.1980/D:14). A method for the reduction of temperature and salinity profile measurements was described in detail.

Information about some of the data and data products available from the ICES Service Hydrographique was given in Doc.C.M.1980/D:17. A plea was made for more data, and more timely delivery of data, for the preparation of some of the marine environmental data products.

5. Discussion. During the discussion various approaches to data storage were considered. The "data explosion" (or perhaps better "data flood") created huge amount of data but the development in computers enabled the storage of them. The problem is the costly retrieval of data from such masses. The data centres are forced to reduce the amount stored in order to be able to work efficiently. Reduction would mean omission of part of the data. But then: what can be left out for one aspect may be important for another.

The solution would seem to be that the data centres store reduced data sets whereas the original data are stored with the originators and can be exchanged bi-laterally.

There is then an obvious need for criteria for what should be stored in the data centres. For STD observations such standards were established by ICES in 1969 and 1971 (see Doc.C.M.1971/C:49), viz.: Minimum standard (Nansen cast equivalent standard), Intermediate optimum standard (99 level standard), Optional optimum standard. However, no such standards have been established for current meter data or tide gauge records. It was suggested that the Working Group on Marine Data Management could deal with this problem. The Group agreed to consider at its next meeting this and other problems raised by the Joint Session.

Such criteria will become increasingly necessary because, as stressed by the author of the Introductory paper (Doc.C.M.1980/D:12), in future oceanographers will more and more have to rely upon data collected by others, i.e., data exchange will continue to increase.

In connexion with the discussion of the need for a commonly agreed accuracy of the data to be stored in the data centres it was stressed that a scientist should not just deliver a number; he should be aware of the inherent inaccuracies and such information should accompany the data.

It was also pointed out that the relative (internal) accuracy of a data set may be high even if the absolute accuracy is low, which means that the data are of much more value to the originator than to others who receive the data mixed up with data from other sources.