International Council for the Exploration of the Sea.

C.M. 1980/D:13
Statistics Committee
ref: Hydrography Committee

THE MIAS OCEANOGRAPHIC DATA BASE - THE DEVELOPMENT OF A GENERAL PURPOSE DATA STRUCTURE FOR SERIAL DATA

by

Meirion T. Jones, Marine Information and Advisory Service (MIAS), Institute of Oceanographic Sciences, Bidston Observatory, Birkenhead, U.K.

Adapted for presentation to the Joint Session of the Hydrography and Statistics Committees entitled "Techniques for coping with the data explosion in marine science" at the 68th Statutory Meeting of ICES, Copenhagen, October 1980.

SUMMARY

In relating to the problem of the "data explosion" in marine sciences, an account is given of how a general purpose data structure was designed and implemented by the Marine Information and Advisory Service of the U.K. for the storage and retrieval of serial environmental data.

****

INTRODUCTION

Although the most obvious effect of the "data explosion" in marine sciences can be examined simply in terms of the sheer rate at which data are being collected, there are a number of associated problems that also deserve serious consideration:-

a) To an increasing extent marine data are required for applications other than those for which they were originally collected. An increasing awareness of the problems of marine pollution and an increasing activity in the exploitation of offshore resources have contributed significantly to the secondary usage of data.

b) The number of laboratories collecting data is increasing such that the available data are spread over many more sources than in the past.

c) Over the years large amounts of data have accumulated in the various data collecting laboratories. It is becoming progressively more difficult and time consuming for individual scientists to maintain such collections and to service requests for data from other users. There is also a risk that the essential documentation necessary to qualify such data may become progressively more inaccessible.

d) There is an increasing tendency towards the study of the marine environment on a global scale and an increasing interest in climatic research. Such trends are likely to increase the demands for the banking and exchange of data.

e) Not only are a wider range of data being collected but an increasing demand exists for interrelating such data in multi-disciplinary studies.

The national oceanographic data centres have a significant role to play in providing solutions to many of the above problems, particularly through:

a) the compilation of inventories documenting the existence of data.

b) the creation, development and maintenance of appropriate data banks.

c) the provision of services to ensure that data can be made readily available to secondary users in a form suited to their needs.

d) the collaboration with other data centres in the exchange of data, both at a national and international level.

One must however be wary of banking and exchanging data simply because it has been collected. Within MIAS for example, all data are assessed according to three basic criteria before they are considered for banking:-

a) The data have been subject to an adequate level of quality control at the data collection stage (encompassing both the original acquisition of the data and their subsequent processing in the laboratory) - including the application of calibration data, the conversion to oceanographic units, the editing out or flagging of erroneous data and the identification of instrument/recorder malfunctions.

b) The data are accompanied by an adequate level of documentation for them to be of value to the secondary user without recourse to the data originator - where possible, such documentation should include information not only on the known limitations of the data, but also on the manner in which they were collected, processed and quality controlled.

c) There is an identifiable need or reason for banking the data.

Data are also assessed according to the effort required to load them onto the data bank relative to the potential value of the data for future use. Whenever possible, data are not considered for banking until they have been analysed or worked up by the data originator, and prior to banking they are screened against the presence of obviously erroneous values.

Although MIAS has concentrated its initial data banking activities on certain specific types of data (instrumentally collected wave data, moored current meter data, sea level data and water bottle data) it was recognised at the outset that, as it became established, the data bank would be required to expand to cater for an ever increasing range of data. It was also recognised that the potential of the data bank to expand to meet these needs was critically dependent on the design of a suitable data structure. The manner in which such a structure was developed forms the subject of the attached paper - certain key characteristics of the structure are worth noting:-

a) Integration - all types of data are integrated within the same structure without the need to duplicate data in order to maintain important data collecting relationships - thus a given series can be associated equally well with other series from the same project, from the same mooring, at the same fixed station, in the same geographic area, of the same type, etc.

b) Generalised - it caters for series containing any combination of measured parameters.

c) Self Documenting - all plain language documentation appropriate to the data is stored with the data.

d) Self-contained Inventory - integrated within the structure is a readily accessible inventory summarising the contents of the data bank - this is achieved without duplication of data.

e) Capacity for growth - it has considerable capacity for growth, both in terms of the number of series it will hold and the volume of individual series.

f) Flexible - it is equally effective in dealing with short series as it is with long series.

The self defining aspects within the structure have much in common with the standard international exchange format for oceanographic data, GF3 - indeed the structure probably represents the closest one is likely to get to implementing GF3 within a data banking environment.

(The attached paper appeared in 'Data base achievements' edited by G.J. Baker, London: A.P. Publications Ltd. and The British Computer Society 1979. The author is grateful to the publishers for permission to reproduce it).

THE MIAS OCEANOGRAPHIC DATABASE – AN INTEGRATED DATABASE/DATA DICTIONARY SYSTEM

by Meirion T. Jones and Trevor Sankey,
Marine Information and Advisory Service (MIAS), Institute of Oceanographic Sciences, U.K.

Oceanographic data is a valuable resource and MIAS has a national responsibility for banking this data so that it is readily accessible to potential users. The nature of oceanographic data and the requirements for the National Oceanographic Data Bank are described. It is shown how a CODASYL type database structure can be evolved to meet these requirements and to provide a generalized and flexible system.

THE ROLE OF MIAS

The Institute of Oceanographic Sciences (IOS), a component body of the Natural Environment Research Council (NERC), undertakes basic research into the seas and oceans. The Institute works in the fields of physical and chemical oceanography, deep ocean biology, and marine geology and geophysics. Much of this work provides essential information for those engaged in practical applications concerned with the marine environment. For example, IOS studies of storm surges are relevant to engineers responsible for flood control, studies of wave characteristics are relevant to the design of offshore platforms or the possible harnessing of wave energy, and studies of ocean circulation assist in the development of safe solutions to the problems of waste disposal.

The Marine Information and Advisory Service (MIAS) was set up in 1976, within IOS, with the aim of making oceanographic data, information and advice readily available to industry, research workers, and local central government departments. There are two aspects to the MIAS Service:

* An Advisory and Enquiry Service which provides information and advice drawn mainly from scientific knowledge and expertise available within IOS.
* A Data Banking Service responsible for setting up and operating a National Oceanographic Data Bank (NODB), and for providing data and information as required by the Advisory and Enquiry Service.

MIAS currently handles about 1000 requests per annum.

The Data Banking Service is located in the IOS laboratory at Bidston, Birkenhead on the same site as the Institute's central computer – a Honeywell 66/20. MIAS has international responsibilities as the U.K. National Oceanographic Data Centre and actively exchanges data, under guidelines established by the Intergovernmental Oceanographic Commission (IOC) of UNESCO, with other national data centres

and with the World Data Centres. MIAS is recognized by IOC as the international data centre for instrumentally measured wave data collected on a worldwide basis.

WHY AN NODB IS NEEDED

In the U.K. alone the annual expenditure on oceanographic research is measured in tens of millions of pounds, worldwide the figure is several hundred million pounds. A significant part of this expenditure arises from the cost of taking measurements at sea. For example, it costs of the order of £5,000 per day to operate an ocean going research ship. Despite the large amounts of money and effort invested in such data collection, the density in time and space of available data from the oceans in general remains low. Furthermore, skilled scientific teams and their equipment are a scarce resource, and it is important that what data is available is made accessible to as many users as possible.

Oceanographic data is valuable not only to the primary user responsible for its original collection, but also to a wide range of secondary users. These secondary users include other scientists, engineers, certification societies, design specialists, marine survey teams, etc. For example, a scientist undertaking a statistical study of the ocean temperature distribution over a particular area may wish to combine data from many different sources. An engineer building an offshore platform, e.g. an oil rig, will design it to withstand the wave conditions to be expected during its working life. Access to wave data is required to derive the design criteria he will use.

If marine data were acquired by MIAS solely as and when requests were received, long delays would be incurred in the servicing of each request. Time is required not only to locate and acquire the data, but also to check its quality, to reduce it to a common format, and to reconcile different methods of data collection and reduction. Such processes are far better conducted on a routine basis so that when requests are received they can be serviced both quickly and efficiently.

Oceanographic measurements are invariably not repeatable – observations of dynamic phenomena are irreplaceable as a record of the conditions existing at the time they were taken. The creation of a centralized national oceanographic data bank not only improves the availability of data to the industrial and scientific user, but also ensures that the data is protected for long term use. Such a bank provides a basis for establishing statistical properties such as the trends, averages and extremes of oceanographic variables, and should prove invaluable for synoptic and climatological studies, whether they be purely descriptive or based on theoretical models.

Although a few types of oceanographic data have been banked nationally or internationally, much data still exists in organized laboratory collections or in the hands of individual researchers. The international marine science community has been concerned over a number of years to improve the banking and exchange of oceanographic data. IOS has had a longstanding interest in this subject, particularly since 1970 and, in setting up MIAS, has made resources available to implement the ideas that had evolved within the Institute.

THE NATURE OF THE DATA

Initially, MIAS is concentrating on banking physical oceanographic data, and in particular, instrumentally measured wave data, moored current meter data and

Nansen cast temperature, salinity and chemical data. These types of data, like most other types of oceanographic data, have certain characteristics in common, although a variety of different instruments and experimental techniques may be used in their collection. Oceanographic data is normally collected and stored in the form of individual data series, each consisting of an arbitrary number of data cycles and a series header containing information associated with the series as a whole. At a higher level the series themselves may be related in various ways. The remainder of this section explains these concepts in detail.

Series and Data Cycles

In most methods of oceanographic data collection, measurements are taken, either simultaneously or in rapid sequence, of a number of different oceanographic parameters. For example, a recording current meter may at a given instant in time measure current speed, current direction and temperature. A Nansen cast bottle when tripped reverses thermometers to register a temperature value (and possibly a pressure value), and also traps a sample of water on which salinity and other chemical determinations can be made. An observer may record a number of different meteorological parameters for a given hour, some read from instruments and others observed visually. In each case the same set of measurements is made repeatedly at specific intervals in space and/or time. The data collected during one repetition of the basic measuring sequence is termed a *data cycle;* the repetition of this cycle builds up into what is collectively termed a *data series.*

Each data cycle will normally contain one value for each of the parameters sampled. Most parameters have numeric values. In a given data series, successive data cycles contain repeated measurements of the same set of parameters. A given parameter set may apply to many different series, but to describe all data series, many different parameter sets, each one a different combination of measured parameters, will be needed. The number of different oceanographic parameters that are measured at present runs into hundreds. In data currently held by MIAS the number of parameters in individual data cycles ranges from 3 to 14, although the majority of the data is held in data cycles with no more than 5 parameters.

Examples of data series/data cycles are:

Data Series	*Typical data cycle*
XBT or MBT dip	depth or pressure, temperature
Nansen cast	depth, temperature, salinity, oxygen
STD/CTD cast	depth or pressure, temperature, salinity
Surface CTD traverse	time, latitude, longitude, temperature, salinity
Geophysical traverse	time, latitude, longitude, water depth, total magnetic field
Recording current meter	time. current speed, current direction
Lagrangian float track	time, latitude, longitude
Fixed wave recorder	time, significant wave height, mean zero crossing period

The main characteristic of a data series is that it links a set of data cycles containing the same set of parameters measured in the same way. This internal

consistency is not only invaluable for studies of the detailed variability of individual parameters, but also means that almost all the qualifying information associated with the actual data is common to the series as a whole and can therefore be held in a series header.

The number of cycles in a series can vary widely from less than 10 to more than 1 million. At present, almost all MIAS data holdings are in series with less than 15,000 cycles. In many cases the length of the series is set naturally. For example, a recording current meter is deployed for a given period and then recovered; the resulting record constitutes a series. A Nansen cast uses a specific number of bottles at a given ocean station. Meteorological observations may cease at the end of a research cruise. In cases where data is recorded for an indefinite period, such as synoptic observations from a coastal location or an oil production platform, the data is usually split into series either by calendar period (i.e. data for each month or each year are treated as separate series), or whenever the instrumentation or method of measurement changes.

Variation in Space and Time

In general, each data cycle contains measurements made at a particular point in time, at a particular geographical location, specified by latitude and longitude, and at a particular depth below the sea surface (or above the sea floor). For a given type of series, one or more of these space-time coordinates will vary, while the others stay effectively constant. The varying space-time coordinates form the key fields for the data cycle or, in scientific parlance, the independent variables.

For almost all series there is a time interval between successive data cycles. If the instrument moves in space during the recording of a series, it is the *space and time scales of the phenomena being measured* that determine whether it is space or time or both that are the key variables. For example, if the time taken to lower an instrument recording temperature through the water column is much less than the time taken for significant temperature changes to occur in that column, then the series can be regarded as a snapshot and the time taken to record it can be neglected. A stationary ship may drift while taking a series of wave measurements in the open ocean without moving to an area where the wave regime is significantly different. In this case, the series can be ascribed to a fixed position. However, the same change in position might be significant for measurements of a parameter varying with a shorter space scale, for example, ocean depth.

Fig. 1 classifies the different types of measurement by their space time characteristics.

Series Type	Methods	Examples	Space-Time characteristics		
			Time	Position	Depth
Time Series	Instrument attached to or suspended at constant depth from a stationary ship, buoy, oil rig, etc.	Shipborne Wave Recorder	Varying	Fixed	Fixed
	Instrument mounted on or moored to the sea floor	Moored current meter, Offshore tide Gauge			
Depth Series	Instrument lowered from stationary ship, oil rig, etc.	Nansen cast, CTD probe	Fixed or varying	Fixed	Varying
	Free falling probe	Sanford E-M profiler, XBT			
Traverse Series	Instrument attached to or towed at constant depth from moving ship	Sea Surface Temperature, Echosounder, Gravimeter	Fixed or varying	Varying	Fixed
	Free floating instrument at constant depth	Swallow float, Drogue			
3-D Traverse Series	Instrument towed at varying depth from moving ship	Batfish	Fixed or varying	Varying	Varying
	Free floating instrument at varying depth	Fully Lagrangian float			

Figure 1. Space-time Variation of Series Data

Series Header Information

The series header consists of the information needed to select, and to present and interpret, the contents of the data series. This information falls into several categories.

* *Identification* — The institution responsible for the collection of the data, reference numbers assigned to the series, etc. The type of data contained in the series, and the type of instrument and instrument mounting used, are also included in this category as they are commonly used to identify series.

* *Location in Space/Time* — The value of each space-time coordinate that is effectively fixed for the series, and the range of those that vary, i.e. the key variables.

* *Qualifying Information* — Documentation of the methods used to collect and process the data, and information on any known limitations of the data. The nature of this information is described in detail in succeeding sub sections.

* *Associated Information* — Additional parameters may be measured which remain effectively constant for the series as a whole, e.g. the surface wind speed or the wave height may be measured at the time of a Nansen cast series. Such values may need to be stored with the series although they do not form part of the data cycle.

Note that the form and amount of this information varies from series to series, and that numeric parameter values, fixed length alphanumeric codes and plain language comments may be included. Some of the information may be specific to a particular series, while other details may be common to a group of series.

Documentation of methods

Inevitably, the methods of oceanographic data collection are tailored to meet the specific aims of the individual collector using the resources available to him. Both the aims and resources vary greatly as do the techniques of measurement, many of which are still under development. These techniques rarely follow precise standards. Although the collector may take the methods he uses for granted and not formally document them, this information is essential for the actual data in the series to have true meaning. Any representation of the data must include this information before it can be regarded as self contained.

The required documentation includes details of the instrument and instrument mounting used, and the methods of data collection (including sampling characteristics), instrument calibration, quality control and processing (including filtering, averaging and compression of data).

Data Quality

There are many potential sources of error in oceanographic data. Most data is derived either directly or indirectly from some form of instrumentation. Instrumental output will usually require the application of calibration results, and it may be necessary to apply corrections based on the performance of the instrument in its operating environment. More severe data problems due to malfunctioning of instrument sensors are sometimes encountered, and problems may also occur with recording systems and their operation, whether they are data logger or mini or micro computer based. Where data is recorded manually, errors are likely both when the data is first written down and in subsequent transcription.

The harshness of the ocean environment, and the vast quantities of data now being collected, have accentuated these problems. Instruments must be used with great care if satisfactory performance is to be obtained. Instruments cannot always be relied on to perform in the same manner in the real ocean as they do in tests under laboratory conditions. Experience shows that the processing of series to give clean, correct data is a major task, one frequently underestimated.

Documentation on the known limitations of the data forms an important part of the qualifying information associated with the data series. If, for example, it is known that data values for a particular parameter are suspect due to an intermittent fault in the corresponding sensor, then this fact needs to be stored with the data itself.

Distribution of data in Space and Time

An oceanographic data bank can never hope to provide a full representation of the oceans themselves in the manner in which a commercial database can, for example, represent a company's activities. Such a data bank would require accurate, consistent data collected at regular, closely spaced grid points in space and time, with each oceanographic parameter being measured at as many points as are necessary to define its variability. Such coverage is not available even for the most commonly observed parameters.

There are logically many links between individual data cycles in different series. There is an obvious connection between all data taken at the same time and between all data taken at the same depth or the same position. However, due to the nature of the data and, in particular, due to variations in measurement characteristics, such associations are often not meaningful. The series as a whole may be logically associated if they are recorded over the same time period, at the same fixed depth, or in the same geographic location.

Although the volume of available data is potentially very large, its density in space and time is in general extremely low, and its distribution is extremely irregular. Not only are the series themselves distributed irregularly, but different series may be based on different sampling intervals (in addition to variations in other measurement or processing characteristics) – vertical profiles of ocean temperature may have vertical intervals varying from 1mm to 1km, sea surface temperature may be sampled once a second or once a day. However, within the irregular distribution of series, there are systematic groupings of data that reflect the data collecting activities of individual organizations, platforms, cruises or experiments. Such groupings are important, in that they usually provide internally consistent data and thus enable certain types of information to be derived as illustrated in the following examples:-

* Three current meters mounted on the same mooring recording current velocity simultaneously at three different depths provide information on the current shear between the different depths.

* Successive wave data series collected at the same fixed station (e.g. an oil rig) provide information on the variation of wave conditions over a longer period than would be possible from a single series.

* A line of current meter moorings laid across, for example, the Northern North Sea during a given experiment provides information on the total flow of water into and out of the North Sea through that section and its variation with time.

* Wind and wave data recorded simultaneously at a given position provide information on the correlation of wind and wave conditions.

THE REQUIREMENTS OF AN NODB

Data Representation

The last section outlined the nature of the data and emphasized its diversity. A National Oceanographic Data Bank system must have facilities to hold both the data cycle and the series header information economically, yet in a readily accessible form.

It is important that the system is fully generalized. If the system were to be tailored only to a limited number of data types, then any requirement for banking further data types would demand system modifications, thus diverting scarce analyst and programmer effort from work on system enhancement. A generalized system on the other hand will be able to input, store and process additional data types without changing the system. Such a system facilitates multidisciplinary studies where correlations are required between different types of data. These advantages outweigh both the effort needed to convert incoming data from a variety of different representations to the standard generalized form and the overheads of generalized processing.

Data Capacity

The space required for the MIAS database is expected to exceed 400 Mb during 1980. This database will contain over 100,000 individual series. In the longer term MIAS anticipates a database of the order of 1000 Mb, although the amount of data that might potentially be included is even greater.

For most types of data, MIAS holdings will be restricted to those ocean areas that are of particular interest to the U.K. However, the MIAS coverage of wave data will be worldwide, in view of its role as an international centre for this data type.

Data Quality

The variety of potential problems has already been described. While much of the data that MIAS acquires has already been cleaned up by its originator, inevitably some errors slip through. It is part of the data banking task to screen series being loaded to the bank for detectable errors, and, where possible, to resolve any problems arising from the data with its originator. It is also necessary to check that sufficient documentation is stored with the data to enable it to be used without recourse to the originator.

Data Retrieval

The retrieval of data/information from the data bank in response to a specific enquiry needs to be considered as a three stage process:-

a. *Search* the data bank to ascertain what, if any, data exists to meet the needs of the end user. As the specific data required may not exist, it often may be necessary to make a fairly broad search of the contents of the data bank in order to identify that selection of available data that is best fitted to the end user's requirements.

b. *Select* the data from the bank according to the criteria derived in the search. This may either be considered as a physical extraction of data or the creation of a 'virtual' file.

c. *Process and Present* the selected data in a form tailored to the needs of the end user. This may, for example, include some form of statistical summary or analysis or a graphical presentation, often in a highly specialized form. Alternatively, if the end user actually requires the data itself in computer compatible form, the selected series may need to be reformatted onto magnetic tape.

It should be noted that, while many of the enquiries to the data bank will be answerable using a specific number of standard retrieval facilities, there will also be many enquiries, of an ad-hoc nature, for which the required facilities cannot necessarily be predicted in advance.

The number of series in the bank makes it essential that a readily available inventory of its contents is available for the search process. The inventory should be held permanently online, in order that the iterative process, often done in conjunction with the end user, of matching the available data to the enquiry can be concluded in a reasonable time. A job turnround time of not more than half an hour is considered acceptable for inventory search, although the ideal would be a fully interactive facility. For the Select, Process and Present stages a turnround time of less than 24 hours is required. Because of the nature of the scientific data held by MIAS, some care is needed in interpreting retrieved values. At present, therefore, it is not intended in general to make the data area of the bank available to end users. The real need of many of those who seek the aid of MIAS is not for data but for information; in this respect, the data bank is seen as a resource for those providing the information service. Access to the data bank, whether for information or actual data, will therefore normally be through a trained intermediary rather than by the end user himself. The value of this approach, recognized for instance by BENNETT (1977), is that the intermediary becomes an expert both at operating the retrieval software and at understanding the capability and potential of the bank and its contents.

Selection Criteria

The required criteria for selecting data series from the bank fall into a number of categories.

* *Parameter* – Series including a particular parameter, for example, data containing measurements of dissolved oxygen. In some cases the same parameter may be measured in more than one way, for example, wave height may be estimated visually or measured from a wave recorder trace. In these cases a facility is needed to retrieve only measurements made in a specified way.

* *Location* – (space and time) Data are usually required for a particular geographic area, specified by latitude and longitude limits. They may also be required for particular depths and/or times, for example, for a given depth range below the sea surface or above the sea floor, or for a given date range or season of the year.

* *Grouping* – Data may be required, for example, from a particular fixed station such as an oil rig, from a data activity such as a particular cruise of a given ship, or from some other series grouping such as a particular research project.

* *Attributes* – in addition to the use of space-time attributes, data series may need to be selected on the basis of virtually any of the attributes assigned to the series header, e.g. source laboratory or country, instrument and instrument mounting categories, or data category (such as spectral wave data or moored current meter data).

Once specific series have been selected, it may then be necessary to further select on the actual parameter values that are contained in the data cycles themselves, e.g. to select data cycles with significant wave height greater than 6 meters or with sea temperature less than 2.0 deg. C.

THE CHOICE OF DATA STORAGE METHOD

Existing Oceanographic File Formats

A great variety of different formats are currently in use by different researchers and laboratories for the storage of oceanographic data. Among characteristics that categorize these formats are:

* *Generality* – Some formats are designed for one specific type of data, for example, current meter data, while others are very general. The generalization is usually achieved by placing a description of the data cycle in the series header.

* *Sparseness* – Some fixed formats specify many fields that are only used occasionally. For example, a format might specify 15 fields in a data cycle of which only 3 or 4 are commonly used. This problem tends to occur with formats originally derived for punched cards. A similar problem arises in the attempt to specify a comprehensive series header using predefined fields.

* *Storage mode* – Some formats use random files, others sequential files. Some are orientated towards disk storage, others towards magnetic tape. Some store data character by character, others store it in binary form. Some include operating system file control information, others do not. The choice made in designing a format depends on the application. A machine dependent binary format offers processing efficiency and convenience. For widespread data exchange, a machine independent sequential format suitable for use on industry compatible magnetic tape and using a standard character code is essential.

* *Header detail* – The amount of associated information given varies widely. Sometimes much of this information is kept separately in manuscript form. Some formats include space for comments, which may or may not be limited.

There are problems with all non database file formats in a data bank situation. In a collecting laboratory, the quantity of data is relatively small and can be indexed manually. The users are usually familiar with the data, they have a detailed knowledge of the methods used to collect and process it, and they are likely to be aware of its limitations; often they are the people who have worked long hours at sea to do the actual collection. Here file based systems work well. For a data bank, the quantity of data is large and the user may not be aware what data, if any, exists to meet his needs. Flexible facilities for selecting and retrieving data to meet a wide range of needs are required; file based systems, whether they are based on large sequential files or on many small files of any mode, are restrictive. In a data bank, as data is used without recourse to the originator, it is necessary to document the information on collection methods and data limitations and to integrate it with the data itself. File based formats are restrictive in this respect also.

The Choice of a Network Database

An evaluation of the techniques and systems available to assist in the creation of an NODB was carried out in 1974 as part of the selection exercise for a new mainframe computer for IOS. Sample data banks were created using six different database management systems, and the facilities of each system were analysed. At that time, it was concluded that the requirement for a fully generalized storage structure able to support a variety of access paths could best be met by a CODASYL type network database management system. This was one of the factors in the choice of a Honeywell 66/20 machine with Integrated Data Store/I (I-D-S/I) database software.

A further study of the problem was made in 1976, when MIAS was formed and there was a need to set up the software for the NODB quickly. To meet the urgent need to prepare documentation for the database design and the system specification, it was decided to seek external assistance and CACI Inc. International, with their particular experience in database design, were engaged for two short contracts. This study confirmed the decision to use I-D-S/I.

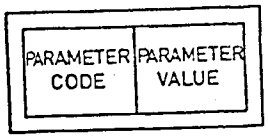
EVOLUTION OF THE DATABASE STRUCTURE

This section describes the evolution of the structure of the MIAS database. A tutorial step by step approach is adopted in order to highlight the concepts involved. The approach starts with the individual data fields and the structure needed to hold the data cycles in a generalized way. The conceptual structure is then extended to represent, first the series header information, and then the required series groupings. Finally, the structure actually implemented by MIAS is presented.

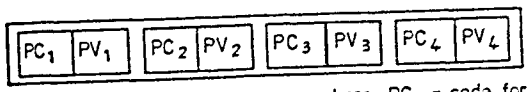
Step 1 – Self describing data cycles (field level)

In order to store individual numeric parameters within a data cycle in a generalized way, some form of data description is necessary. One conceptually simple method is to use self describing fields, where each parameter value is coupled with a parameter code to identify it (Fig. 2a). A program retrieving this field can then interpret the code to determine the parameter to which the numeric value refers.

a) SELF DESCRIBING FIELD



b) SELF DESCRIBING DATA CYCLE - PAIRED COUPLES



where PC_n = code for n^{th} parameter
PV_n = value for n^{th} parameter

c) SELF DESCRIBING DATA SERIES - PAIRED COUPLES

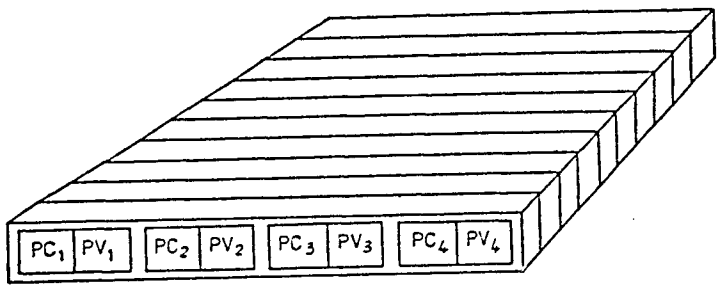
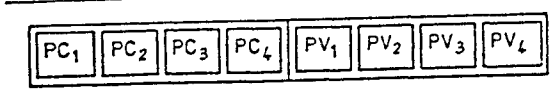


Figure 2. Evolution of Record Structure – Step 1

a) SELF DESCRIBING DATA CYCLE - SEPARATED COUPLES



b) SELF DESCRIBING DATA SERIES - SEPARATED COUPLES

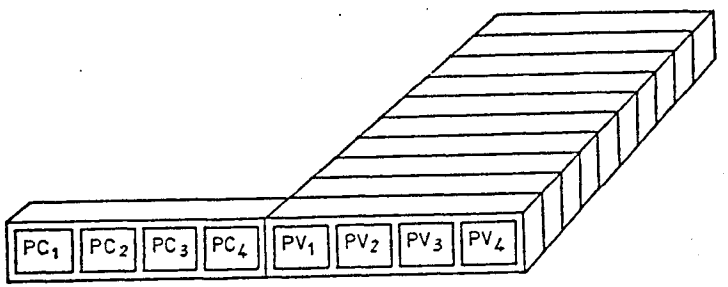


Figure 3. Evolution of Record Structure – Step 2

For example, a numeric value of 12.1 linked to the parameter code signifying temperature will indicate a temperature value of 12.1.

A number of such self describing fields can be placed together to form a self describing data cycle, as shown in Fig. 2b. Many such data cycles can then be built up into a data series, as shown in Fig. 2c.

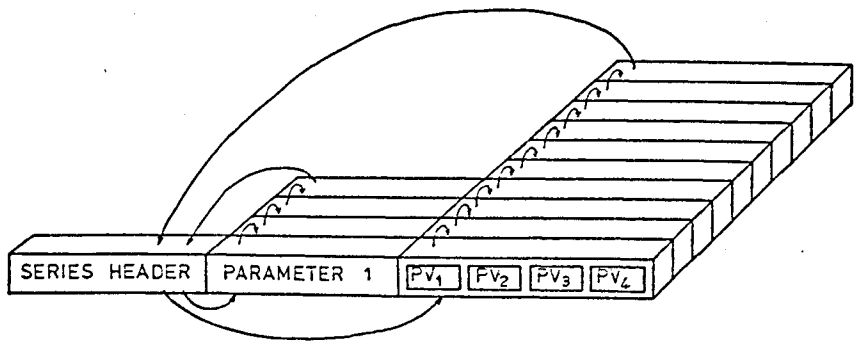
Step 2 – Self describing data cycles (series level)

The evolution of the structure can be continued to the stage shown in Fig. 3a, where the parameter codes are separated from the parameter values, so that the data cycle contains, first the parameter codes in order, followed by the parameter values in corresponding order. For most oceanographic data, each data cycle within a given data series contains values for the same set of parameters. It is unnecessary, therefore, to repeat the set of parameter codes in each data cycle and the series format shown in Fig 3b can be used.

Step 3 – Structured data series

The data series is not complete without its series header information which is represented at this stage as a single record, as shown in Fig. 4a. As the set of para-

a) SINGLE SELF DESCRIBING DATA SERIES



b) BACHMAN DIAGRAM OF FIG. 4a

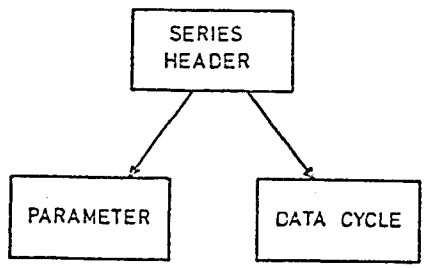


Figure 4. Evolution of Recprd Structure – Step 3

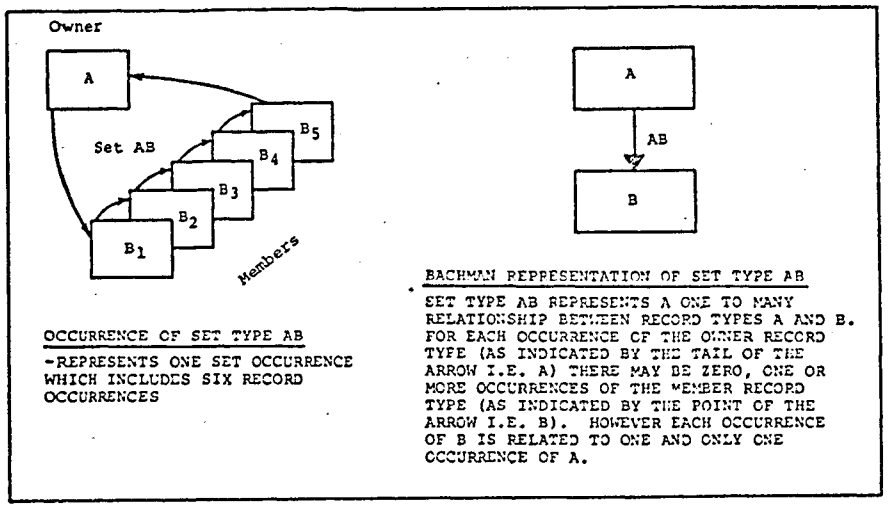


Figure 5. Bachman Data Structure Notation

meter codes is only held once per series it can be represented as a set of individual parameter records each containing additional information about a specific parameter, such as its units and maximum and minimum values.

It is convenient at this point to introduce the network database concept of a set (alternatively termed a co-set or chain). A named set type consists of an owner record type and one or more member record types. An individual set occurrence consists of one occurrence of the owner record type and any number of occurrences (including zero) of the member record types. A set thus supports a one to many logical relationship. The order of the member occurrences within the set may be controlled and used to carry information.

The relationship of the records shown in Fig. 4a can be expressed by two set types, both owned by the series header record. In the first set type the parameter record is the member type, while in the second set type it is the data cycle record that is the member type. As the figure shows, an individual series header is related to a number of parameters and to a number of data cycles. The order of the data cycle records within the one set usually maintains the chronological order in which they were recorded. The order of the parameter records within the other set remains in correspondence with the order of the data fields in each data cycle record. For example, if the third parameter record in the set describes current direction, then the third data field in each data cycle record will contain a value for current direction. This correlation can be used by retrieval programs to identify particular parameters in a given series.

Networks of records and sets can be represented by a shorthand notation known as the Bachman diagram which is explained in Fig. 5. Note that it is convenient to extend the notation to represent many to many relations as, although these cannot be directly implemented, they enable the conceptual structure of the database to be built up concisely. Fig. 4b shows a Bachman diagram for the structure of Fig. 4a. For clarity the set names have been omitted from Figs. 4b to 9.

Fig. 4a includes pointers that illustrate one method of implementing sets. Each pointer would be implemented physically by holding the storage address of the record pointed to in the record from which the pointer stems. It is this explicit maintenance of relations between records that enables complex network structures to be built up, giving a flexibility absent from sequential files where the relations between records are determined implicitly by their placement on storage and are restricted to hierarchical type structures.

Step 4 – The Data Dictionary

The structure of Fig. 4b contains two types of redundancy. On the one hand, many series contain the same set of parameters, for example many current meter series contain data cycles composed of values for the parameters time, current speed and current direction. On the other hand, an individual parameter such as time or current speed may occur in a number of different parameter sets. It is obviously undesirable to repeat all the supplementary information for each parameter for each parameter set it appears in.

The structure shown in Fig. 6 describes each parameter once. The parameter record includes the parameter code, the units in which the parameter value is

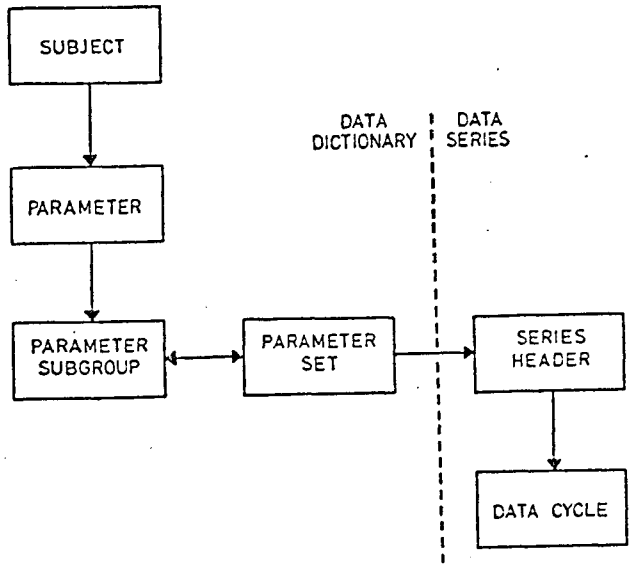


Figure 6. Evolution of Record Structure – Step 4

stored, nominal minimum and maximum values for the parameter, its storage mode and format, and full and abbreviated versions of the parameter title for use in data presentation. The parameters themselves are grouped under subjects, e.g. waves, currents, winds, in order to provide access to the database by subject.

Each set of parameters is defined by a parameter set record which is linked to each of the appropriate parameter records. Note that this is a many to many relationship: each parameter set is linked to many parameters, while each parameter may be linked to many parameter sets. The order in which the parameter records are linked to the parameter set record is again maintained to correspond with the order of data fields in the data cycle records. Each series header record is then linked to the appropriate parameter set record.

In fact the parameter set record is linked, not to the parameter record, but to a parameter sub group record. This supports the requirement, previously outlined, to discriminate between measurements of the same parameter made using fundamentally different methods. The sub group record includes the code and full title of the sub group.

Fig. 6 represents an outline of the data dictionary facility embodied in the MIAS data bank structure.

Step 5 – Series header and document records

The contents of the header of a data series are very much dependent on the type of data but can generally be divided into three parts.

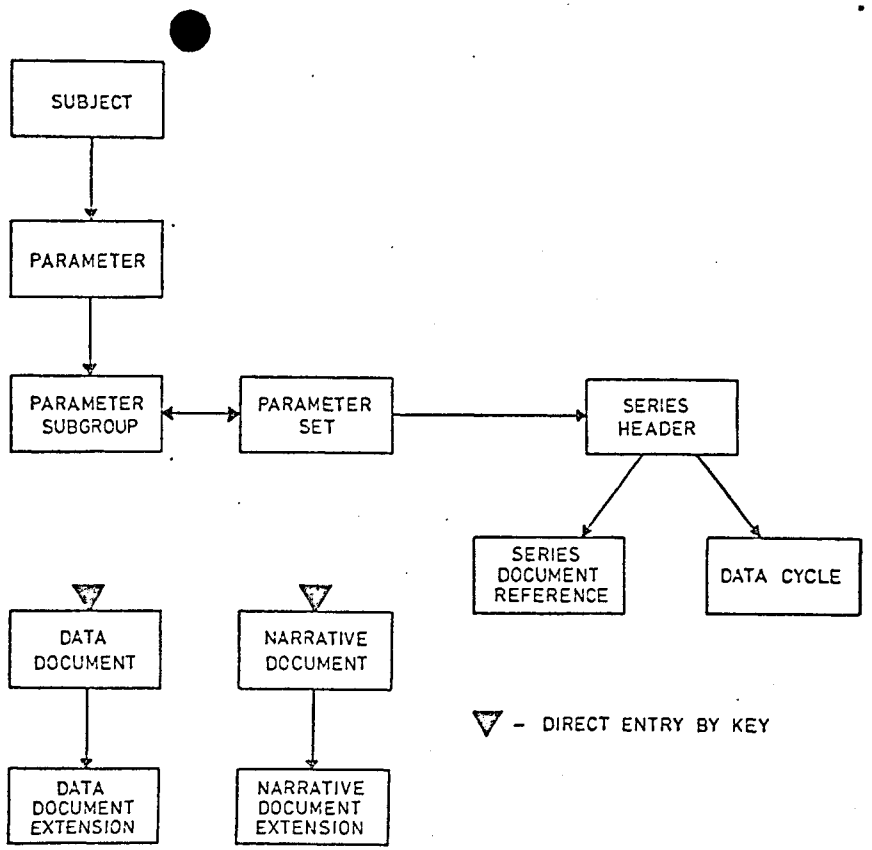


Figure 7. Evolution of Record Structure – Step 5

A) A fixed format part that is independent of the type of data being stored, and contains fields to identify the series and to specify its location in space and time (expressed as ranges or fixed coordinates depending on the type of series).

B) Plain language text containing the documentation applicable to the series.

C) Additional data fields containing any parameter observations associated with the series as a whole, and also any measurement attributes of the data cycle parameters e.g. cross sectional area of a towed biological sampling net. These fields may be needed in conjunction with those in part A for the subsequent selection or processing of the data.

The contents of part A are stored in the series header record itself. However, in preference to reserving a lot of space for optional information within the series header record, document records are provided within the database to accommodate parts B and C. There are two types of document: the narrative document (for part B) which holds one or more lines of text, and the data document (for part C) which

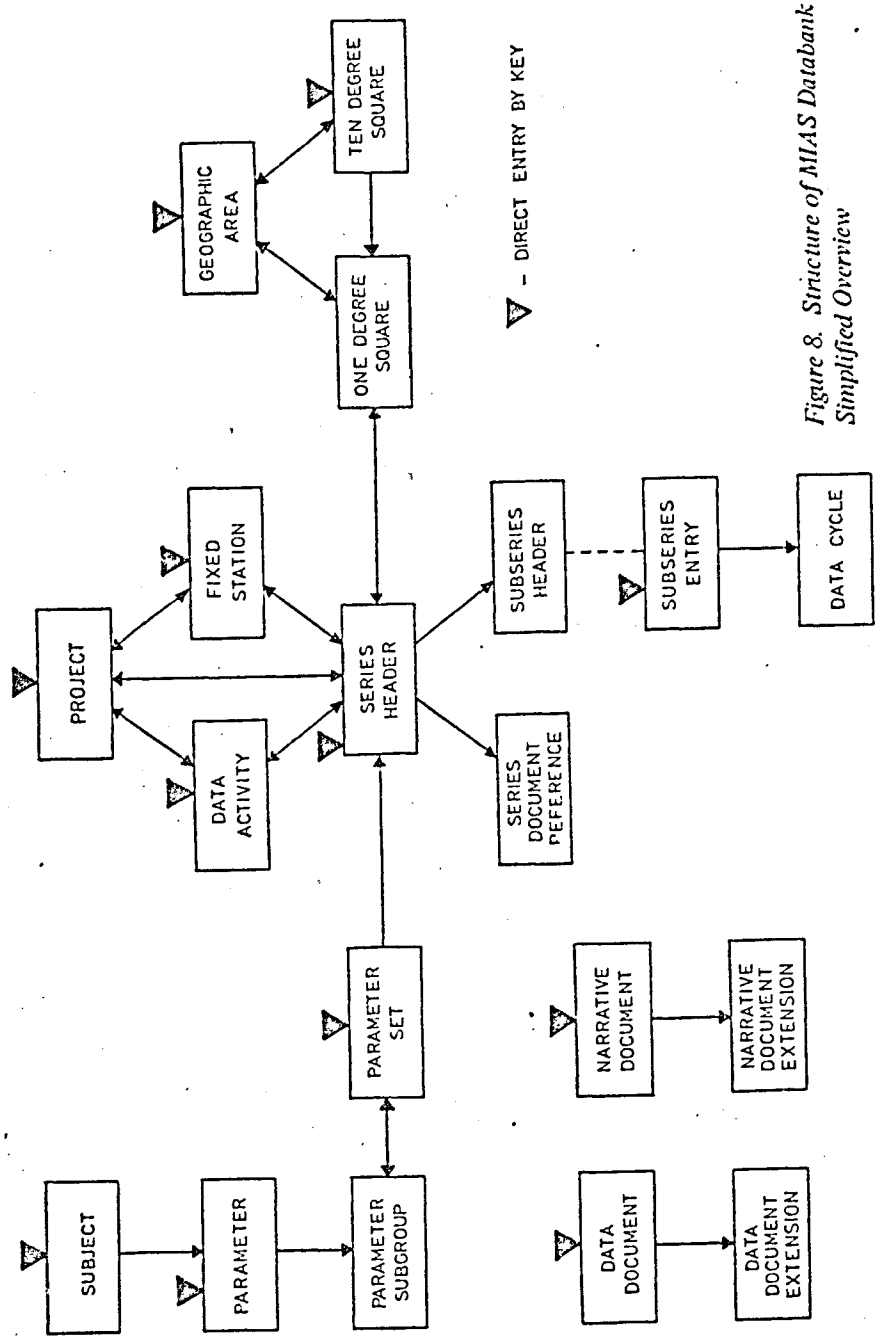


Figure 8. Structure of MIAS Databank – Simplified Overview

consists of a group of self describing couples each consisting of a parameter code followed by a parameter value.

Each document may contain one or more records. The first record in each is a calc record that may be accessed directly using the document reference as a key. If necessary, this record may be extended by linking to it a set of extension records. Each record (calc or extension) holds one line of text or six self describing couples. The self describing couples use the same parameter codes as those used to describe data cycle fields. It follows that a full description of each parameter can be obtained by directly accessing the appropriate parameter record.

Both narrative and data documents are referenced from the series header record by a set of series document reference records (Fig. 7), each holding the key of a document and a flag for the class of information contained in the document (distinguishing, for example, data quality warnings from associated environmental data). Each document can be referenced by one or many series headers.

A limited extension facility is provided for certain other records in the database, namely project, data activity, fixed station, geographic area, parameter, parameter sub group, parameter set and sub series header records, by reserving fields in each record to hold the reference number of one document of each type.

Step 6 – Grouping of series

The structure is required to maintain the relationship of systematically collected groups of series. Separate record types are included for three classes of group:

a. *Fixed Station* – A specific location where many data series are collected as a result of deliberate planning. Examples are a location to which a ship regularly returns to take measurements, a location monitored over a considerable period by an anchored platform such as a weathership, a light vessel or an oil rig, or a location where a number of different series are taken simultaneously over some period, for example, by a data buoy or a heavily instrumented mooring. A fixed station is identified by nominal geographic coordinates of either a fixed point or a small area. How close a measurement must be to the nominal position for the data series to belong to the fixed station depends on the type of data collected and the nature of the surrounding area.

b. *Data Activity* – An identifiable programme of data collection conducted by a given laboratory from a specific platform during a limited period of time typically, but not necessarily, of the order of a month. Examples are the cruise of a ship or a three month observing period at an oil rig.

c. *Project* – Any other logical grouping of data, large or small, that it is considered worth supporting. Both individual data series and the entire data holdings for fixed stations and data activities may be linked to a project. A particular use of the project record is to link together data from large scale experiments that cover more than one platform, laboratory and/or time period.

The representation of these groupings involves many to many relationships between series and fixed stations, data activities or projects, and also between fixed stations or data activities and projects as shown in Fig. 8.

Step 7 – Structure entry points for data access

In the structure so far derived, entry points to the database are provided through the subject, parameter, parameter set, project, data activity, fixed station and series header records. All these records are defined as calc records and can be accessed directly by specifying the appropriate identifier. Entry by geographic coordinates is also required and this is supported by introducing geographic area, ten degree square and one degree square records as shown in Fig. 8. Geographic selection can be specified directly as latitude and longitude values or ranges, with data access provided through the ten and one degree square records. Alternatively, named geographic areas (e.g. the Celtic sea), defined in terms of ten and one degree squares, can be accessed. Overlapping of defined areas is permitted but there is no provision to subdivide one degree squares.

Each data series is normally linked to all the one degree squares for which it contains data. Time and depth series taken at a fixed position will of course be linked to only a single one degree square, while traverse series may be linked to many.

Step 8 – Series subsetting

In order to facilitate access to, and indexing of, data in long data series the concept of the sub series is introduced. The sub series is the smallest grouping in which the actual data cycles can be accessed. The sub series header record acts as an index record for a group of data cycles and contains only the logical reference to where the group is stored and certain indexing information. The definition of sub series is very flexible, but the normal practice is to start a new sub series each time either the month or one degree square of observation changes at the data cycle level. This enables data to be selected from a given one degree square by month of observation, by first selecting data series headers with the appropriate space-time ranges and then checking the index variables in each subordinate sub series record for the required combination.

Step 9 – Offline data cycle storage

In the structure of Fig. 8 the actual data cycles constitute a high proportion of the total storage requirement of the data bank. Moreover, the data cycles from any particular series will, in general, be accessed by only a small proportion of the total number of enquiries answered from the databank. The preliminary question as to what data is available to meet an enquiry can usually be answered without any reference to the actual data cycles. Indeed, the structure down to sub series header level provides an inbuilt inventory of what data is available in the bank.

It is therefore convenient to separate the physical storage of the data cycles from that of the rest of the structure. The data cycles for a given data series are linked sequentially to a sub series entry record. The sub series entry record is reached by calc access using the series and sub series reference numbers as keys. The link between sub series header and the corresponding sub series entry record is thus a logical one as symbolized by the dotted line in Fig. 8. To allow storage of data cycles in a number of database realms (areas in I-D-S/I) the realm number is also held in the sub series header record.

This structure allows the header information for all series to be held permanently on-line. However, only the data cycle realms containing the data cycles required to answer the current enquiry need be loaded. It follows that the data cycle realms may be stored on removeable disk packs and that their total volume is not limited by the disk drive capacity of the computer system being used.

Step 10 – The finished structure

The actual data bank structure implemented by MIAS is shown in Fig. 9. Essentially the structure is the same as that of Fig. 8 except that:

a. Linker records have been introduced to support the many to many relationships of Fig. 8.

b. Six different data cycle records are defined to hold different numbers of parameters in the range 3 to 14 – only one of these is used for any given data series. Additional data cycle records may be added as and when data cycles with more than 14 parameters are to be stored.

c. Primary records have been added to allow sequential access to most of the major access entities on the data bank. Sequential access to all ten degree squares is provided within the structure by defining a geographic area "World" to which all ten degree square records are linked.

Developments a. and b. above were introduced in order that the structure could be implemented using the facilities available in I-D-S/I. Codasyl type databases do not support many to many relationships directly, and variable length records are not supported by the I-D-S/I DBMS used by MIAS.

It should be noted that each individual parameter value within a data cycle record has associated with it a one character flag for the purpose of qualifying the value, e.g. to flag that the parameter value is suspect or to flag that it is missing from that particular data cycle. This facility enables series, where one parameter is sampled less frequently than other parameters, to be expressed in the form of unstructured data cycles, e.g. if a given parameter is only measured in alternate data cycles, then its value in every other cycle is flagged as absent.

DISCUSSION ON THE STRUCTURE

Properties of the Structure

The structure exhibits three properties that are fundamental to the development and maintenance of an effective NODB:

Generalized and flexible – the structure is just as applicable to wave data as it is to current data, or to sea level data, or indeed to any series of marine or other environmental data that can be expressed in terms of unstructured data cycles each containing values for the same set of parameters. The structure provides full generality at the data cycle level, and sufficient flexibility is provided at the series header level, through the use of narrative and data documents, to ensure that full qualification and documentation of the data can be stored as part of the data series. The document records also allow useful flexibility to be provided at other points in the structure.

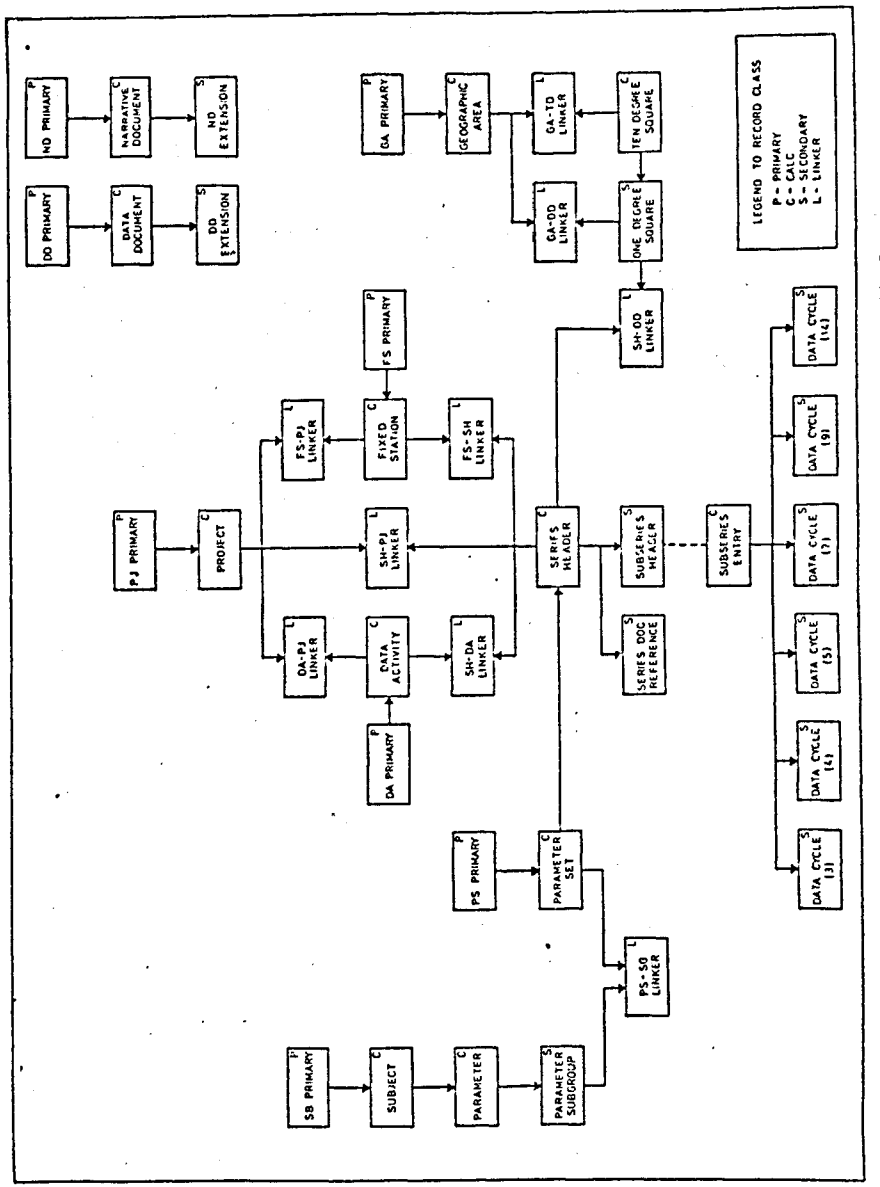


Figure 9. Detailed Structure of MIAS Databank Version – August 1978

Integrated – In addition to their relationships to oceanographic disciplines and to geographic areas, series may be related in the project, cruise or fixed station groups in which they were collected. Many different viewpoints are thus integrated within the same system. The bank is just as much a wave databank as, for example, it is a data bank containing wind, wave and current data from an oil rig in the North Sea, and a multidisciplinary data bank for the Celtic Sea. This integration is achieved with minimal redundancy.

Inbuilt Inventory – The portion of the structure down to the sub series header level, that provides an inventory of the contents of the data bank, is stored as a single integrated area independently of the data cycles themselves. This enables all data selections (except those based on the value of particular parameters in the data cycle records) to be fully resolved without requiring access to any of the data cycle areas. This is extremely important because, in the formulation of any enquiry to the data bank, it is essential first to ascertain what data is available to provide an answer. Only then can it be decided what data should be selected and how it can best be processed and presented to meet the needs of the enquirer.

The authors are not aware of any similar system using a Codasyl database.

The Structure in Practice

The structure described is that in use since the summer of 1978. This differs only in one major way from the first complete database description which was produced in November 1976. The problem in the earlier versions was the geographic access path which was linked in at the sub series level. This caused some anomalies which the current structure resolves.

Some aspects of the structure are more well proven than others. The basic representation of series and sub series has proved very sound, and the data dictionary portion of the structure works well. The structure for access by geographic postion is also effective in its revised form, and the narrative document facility has been found most useful. Features attracting less use have been the facility to group ten and one degree squares into geographic areas, and the data documents which are somewhat cumbersome for heavy use with their self describing data fields. The worth of data activity, fixed station and project records is awaiting evaluation, as they have not yet been used extensively.

It should be noted that the techniques used in this structure are not limited to a scientific situation, but might usefully be applied in normal commercial data processing.

Implementation Aspects

The current implementation uses I-D-S/I. Run time access from batch Fortran programs is provided using an Interface written by Honeywell according to IOS specifications.

Honeywell now market I-D-S/II as part of the ASCII based DM IV data management package. This system is a Codasyl implementation and now includes embedded Fortran DML facilities. IOS may convert to this product in the longer term when a clear reason for doing so emerges. At present, the additional costs of conversion,

software charges and the extra machine resources needed outweigh other considerations. An I-D-S/II version of the database would be very similar in concept to the present one. In detail both the disk format and the Fortran DML are different; conversion would require both program changes and an unload, reformat, reload operation on the data bank, equivalent to a considerable restructuring.

Restructuring of the data may become necessary in the future for a number of reasons:

* The inventory area of the base is overloaded.
* A data area becomes full and must be split or enlarged.
* A change in the data bank structure is needed.

The solution used will depend on the problem. While data volumes are small, data will simply be reloaded. This will soon become impossible, because of the data volume, and tailored restructuring programs will have to be written. Restructuring is made much easier by the time scales of the application. Provided the current version of the bank is available to meet enquiries, updating could be stopped for up to a month while a restructured version of the bank is being prepared. In emergency, an even longer break in updating could be tolerated.

The use of a DBMS has not caused performance problems. Databank retrieval forms a very small part of the machine load and convenience is more important than efficiency. By permitting only well trained users to access the bank, it is hoped to minimize the problem of queries that consume excessive computer resources. Many of the databank records include counts of the series or cycles stored subordinate to them so that it will be possible to ascertain the approximate time and cost for retrieval and presentation for a given query before the data is actually accessed. It must be remembered that at present the system is not a real time one. The provision of multiple access paths means that the amount of sequential processing is kept small and this compensates for the DBMS overheads. Character I/O in Fortran is notoriously slow and the use of I-D-S with the Fortran interface avoids this bottleneck. For series containing large numbers of data cycles a facility is planned to hold the cycles on sequential files on either disk or magnetic tape, and to reference them through an index record in the inventory area of the database.

Using the Structure

The software components of the system being built by MIAS using the data bank structure of Fig. 9 fall into three general areas. The Conversion domain covers all data handling from when data is first received until the data is ready for loading to the database, including all data screening. The Load domain is responsible for actually loading data onto the bank and for all DBA functions, including data security. The Retrieval domain covers selection and retrieval of data from the data bank and subsequent processing and presentation.

So far the investment in developing the software system has amounted to approximately five people for a period of two years and effort is continuing at this level.

a. *Conversion*

In the early stages of building the NODB it is essential to concentrate on acquiring sufficient data of good quality to provide an effective base from which to service user enquiries. Accordingly, much effort in MIAS has been devoted to the Conversion domain, both for software development and for operational data handling. Currently over half of the staff effort in the MIAS Data Banking Service is being applied in this area. While the conversion system is designed so that the later stages of the processing use common software for all types of data, the initial stages must be modified for each format in which data is received and screening methods also vary with the data type.

b. *Loading*

Only one person is currently working in this area both operating and extending the system as the requirements are relatively stable.

c. *Retrieval*

To give flexibility and to reduce main memory requirements the retrieval function is divided into several stages.

The first stage, which may be required on its own, is to determine what data series are available to meet the needs of a specific enquiry. The criteria for selection are expressed primarily in terms of the access paths provided – series, parameter, location and series grouping. Selection can also be made on the fixed fields of the series header. This stage uses only the inventory portion of the data bank. While the quantity of data loaded is small relatively little effort has been devoted to this stage. Considerable further effort is being planned.

In the second stage, for which the required data areas must be on line, actual data is retrieved from the series on the bank to meet the data selection criertia which specify, within each of the data series, the information required. Retrieval can be limited to named parameters – for example only depth and temperature may be required from a series also containing salinity and oxygen values. Particular value ranges of one parameter can be selected, for example, wave height values can be selected from a series for the known time interval when a storm occurred.

For the third stage, processing and presentation, a large suite of presentation software is available, developed from existing IOS software. This interfaces to the database retrieval stage via an intermediate file format, PXF, which is also used by the Conversion domain as the initial format into which data is converted. An advantage of a generalized system is that the processing and presentation can be applied to any suitable type of data. An X-Y display can be plotted for any two fields using a single program, as all the parameter details are obtained from the data dictionary. In practice, the user requires the presentations that are customary for the type of data he is looking at. Hence new types of data will often require the development of additional presentations.

Comparison with Commercial Database Applications

The data analysis and the database design for the MIAS data bank show a number of characteristics not commonly found in business database applications.

Two primary characteristics are:

* *Non-functional data* – In the commercial world the data being modelled is usually directly related to the administration of the operations of some part of an organization. This means that the database is volatile with a large volume of both updating and retrieval, on a daily basis for the main applications. The MIAS data bank is non-volatile; once data is loaded to the bank it usually remains there indefinitely.
* *Generalization* – The MIAS data bank is required to support a variety of incoming data types and of retrieval access paths. To meet this goal, the data structures and access paths are mapped into a generalized format at an appropriate level of abstraction. While this happens to some extent in commercial databases, in general the level of abstraction is lower.

Some further characteristics stem from these two. In a commerical enterprise there is little problem in defining the limits of the data relevant to the organization; in science, organizational boundaries are of less significance as data collected by different groups or organizations, possibly in different countries may be closely related through a common project or a common area of interest. In other cases, different sets of data collected by the same group may have minimal scientific relationships. These patterns of relationships blur the normal distinction between global and local data models.

In the last few years, increasing emphasis has been given, in the design of commercial real-time applications, to the detailed tuning of database structures on the basis of transaction volumes. The generalized nature of the access requirements for a scientific data bank prevents this approach. The access paths that may be required are limited only by the nature of the data. While the most common access paths are known from previous experience, the use of other paths is not predictable. In science, a problem can only, and will only, be studied when the tool to measure it becomes feasible with the technology available. It follows that the pattern of usage of data access paths will depend on the pattern of provision, and the data bank designer has to choose a suitable compromise between user freedom and implementation expense.

Compromise is also needed in the choice of the level of abstraction. If many specific entities are supported, then the relationships between them will become tangled; if the entities are made very general, more mapping information is needed to interpret them. The same choice is needed at attribute (field) level. In the MIAS data bank structure, the series header, the data cycles and the document records are generalized entities, but, for data groupings, the more specific entities, project, data activity and fixed station, are used. A virtue of generalization is that a generalized structure can be applied to meet needs that were not envisaged when the structure was designed.

It may be argued that the differences described are differences of degree rather than of substance; nonetheless their combined effect is to make the design of a scientific data bank a distinctive task. TAGG (1978), in discussing the data analysis and database design for a quite different scientific data banking application, isolated similar conclusions.

The Future

The present version of the data bank is restricted to the storage of series containing unstructured data cycles, i.e. data cycles that do not themselves contain repeating groups. However, in time series containing, for example, three hourly one dimensional or directional wave spectra, each individual data cycle will in fact contain a series of spectral estimates, and each spectral estimate may itself exhibit the properties of a data cycle within the main data cycle. Further development of the data bank structure is planned to permit the storage of such structured data cycles.

Parameter descriptions in the data dictionary part of the present structure apply only to fields stored at data cycle level and to fields in data document records. In future it is probable that the description will be extended to fixed fields in the series header, location and data grouping records. This will enable a program retrieving a predefined field, such as "latitude" from a series header record, to be given a full set of attributes for that field for use by generalized processing programs. Provision of this facility will allow the retrieval of "virtual series", series in which each cycle is made up partly of header data and partly of cycle information. For example, to produce a temperature contour map at a particular level from a number of depth series of temperature, a series must be produced composed of the position from the header of each series and the temperature value from the datacycle in that series containing the specified depth value. (In practice, different series will collect data at different depth horizons and most will not have data exactly at the required depth. It follows that an interpolation would be required to derive the required temperature value.)

The inclusion of fixed field attributes will be a valuable extension to the self describing nature of the MIAS data bank structure.

ACKNOWLEDGEMENTS

The authors wish to express their thanks to Mr. G.J. Baker of C.A.C.I. both for his invaluable contribution to the development of the structure of the MIAS Database and also for providing some of the material on which the early parts of the section on the evolution of the database structure were based.

The work described in this paper was funded jointly by the Department of Industry and the Department of Energy.

REFERENCES

BENNETT, J.L. (1977) Expanded roles for information transfer specialists in interactive information management. Rep. RJ 2025, IBM Research Laboratory San Jose, U.S.A. Unpublished Manuscript.

TAGG, R.M. (1978) Data analysis for scientific databases. In: Data analysis for information system design, conference papers 29th June 1978, R.N. MADDISON, editor, British Computer Society, pp RMT I – RMT 4.