ICES STATUTORY MEETING 1993

ICES C.M. 1993/D:23
Sess.T

# AN INTEGRATED DATABASE FOR MARINE RESEARCH

by

S. Floen, H. Gjøsæter, R. Korneliussen, H. Sagen, P. Thorvaldsen and V. Wennevik

Institute of Marine Research
P.O. Box 1870, N5024 Bergen, Norway

## Abstract

In 1992 an integrated model for marine research data was developed at the Institute of Marine Research. Traditionally, the marine research disciplines collect data from their own fields, normally used in a limited context. In many research programs there is a strong need to cross-relate data from various sources. To accomplish this it is essential to establish an integrated database, with a uniform data representation and a structure that is flexible enough to represent all kinds of marine data. A centralized database increases the availability, avoids redundancy and enforces access control and protection against loss of data.

The model is based on a generic data structure where “platform”, “operation”, “object” and “measurement” are central entities. In an operation (e.g. trawling) on a platform (e.g. ship), series of measurements (e.g. lengths) of different objects (e.g. fish) are gathered. Data can be classified according to different systems (e.g. taxonomy, regions and methods). The database is implemented in the Ingres relational database system running on HP9000/755 computers with Unix operating system and connected to the institute’s local network and Internet.

Tests have been carried out to evaluate functionality and performance. While the test results are acceptable, minor immediate modifications are needed. The database will be subject to further development in the future. Currently, work is in progress on data quality assurance, inserting data into the database and adapting software applications. The database is planned to be in regular use at the end of 1993.

# 1 Introduction

On the international scale, the Institute of Marine Research (IMR) is a major collector of marine data. The institute registers and possesses large amounts of data which form an important part of the national and international basis for research and consultancy in fisheries and environmental management issues.

For many decades, the handling of data at the IMR was an entirely manual process. The measurements made: fish length, water salinity, or weight of a plankton sample were written on specially designed forms. The resulting piles of paper formed the Institutes database. The database management system (DBMS) was totally manual, and the relevant papers had to be retrieved from the filing cabinets. Then people had to read the data from the forms, undertaking any necessary calculations either by hand, by mechanical computing devices or, from the 1970s, by electronic calculators.

In the 1970s, the first computers were installed at the IMR, and these were immediately utilized for storing (on plain ASCII files) and manipulating (by means of FORTRAN programs) data. The next decade was characterized by the emergence of the personal computer. One of the implications were that each scientist used their own databases.

The new trend in the 1980s was to look for relations between various types of observations. And suddenly the heterogeneity of the databases at the IMR was experienced as a major drawback for integration of work in the different fields of science.

At the start of the 1990s the dissatisfaction with the situation gave birth to a group, consisting of the authors of the present paper, with terms of reference to develop an integrated model for all, or as much as practical, of the Institutes data, and to implement this model in a DBMS accessible throughout the Institute via the Unix-based network which form the basic resources of IMR's computing resources today. This work, which is presented here, forms a solid basis for an integrated information system at the IMR.

# 2 Data Sources

The activities at IMR are divided on three departments, Department of Marine Resources, Department of Marine Environment and Department of Aquaculture. The data collected are of different types and nature, and traditionally they are local to the departments.

## 2.1 Marine Resources

The Department of Marine Resources does research on the living resources in the sea, i.e. the stocks of fish, shellfish and marine mammals. The research concentrates on studies on stock sizes, age and growth, feeding, and migrations. Studies on how stocks interact and affect each others growth and recruitment is central. Gear technology also forms part of this departments responsibility. The bulk of the data is sampled during surveys with IMR's research vessels.

The work can be divided into the following categories:
- Reproduction and recruitment
- Ecology and multispecies modelling
- Stock size estimation and stock structure
- Resource appraisal for management purposes
- Environmental quality and fish health
- Gear technology and fish ethology

The main types of data are:
- Biological data (length, weight, age etc.)

- Acoustic data (acoustic density estimates, target strength data etc.)
- Marking and recapture data
- Stomach analysis data
- Gear technology and fish behaviour data

## 2.2 Marine Environment

The Department of Marine Environment does research on physical, chemical and biological parameters concerning the management of the marine environment in the ocean and in Norway's costal waters.

Research programmes in 1993 are:

- Ocean climate
- Marine ecology
- Reproduction and recruitment
- Monitoring and assessment of the marine environment
- Radioactivity and pollution

For the purpose of studying the environment, life and interaction of marine species, the following types of data are collected:

Data sampled at specific stations:

| | |
|---|---|
| • Hydrographic data: | Salinity, temperature, depth, etc. |
| • Chemical and biological data: | Nutrients, phyto- and, zooplankton, fish egg and larvae |
| • Radioactivity and pollution: | Cesium 137, various pollutants |

Continuous data:

| | |
|---|---|
| • Thermograph data: | Salinity, temperature, depth |
| • ADCP data: | Acoustic Doppler Current Profiler |
| • ARGOS data: | Drifting buoys |
| • Current data: | Speed, direction, salinity, temperature |
| • Meteorology | |
| • Fluorescence | |

## 2.3 Aquaculture

At the Department of Aquaculture, the research is focused on developing better methods for the commercial production of freshwater and marine organisms in aquaculture. The department also contains units that work with diseases in fish and shellfish and population genetics of wild and reared populations of several fish and crustacean species.

Most of the research conducted at the department is experiment oriented, studying variables in closed, controlled ecosystems. Such experimental data does not lend itself readily to inclusion in the chosen database model. We have therefore chosen to include in the database only those data that comply with the general model of the database, e.g. observations of natural variables in open ecosystems. The data that fall into this category at the Department of Aquaculture are observations on the natural variability of genetic markers in wild populations of fish and crustaceans, analysed by electrophoresis.

# 3 The data model

In many marine research programs there is a strong, growing need to use data from different sources, out of the departments where the data is collected. To meet the requirements of access to all IMR's scientific data in wanted combinations, a conceptual, unifying model of the institute's data was developed.

To get an overview of the different data entities and the relationships between them, a simplification of the Entity-Relationship model technique was used. During the iterative process of refining the conceptual data model, it appeared, at an early state, that all time series oriented data had the same basic structure. It was therefore decided to use a general structure, capable of representing all kinds of marine research data as the main principle for the data base. The generality is achieved by using a generic structure, where "platform", "operation", "object" and "measurement" are the main entities. Data is represented as an operation (e.g. trawling) on a platform (e.g. ship) with series of measurements (e.g. lengths) of different objects (e.g. fish).

A general organization of the database makes it flexible, but also leads to a risk of overloading the DBMS with complexity and inefficiency as a result. For some kinds of data, in particular data collected in large series and amounts, special tables are defined. The special tables supply some of the general tables connected by a common design and use of database keys.

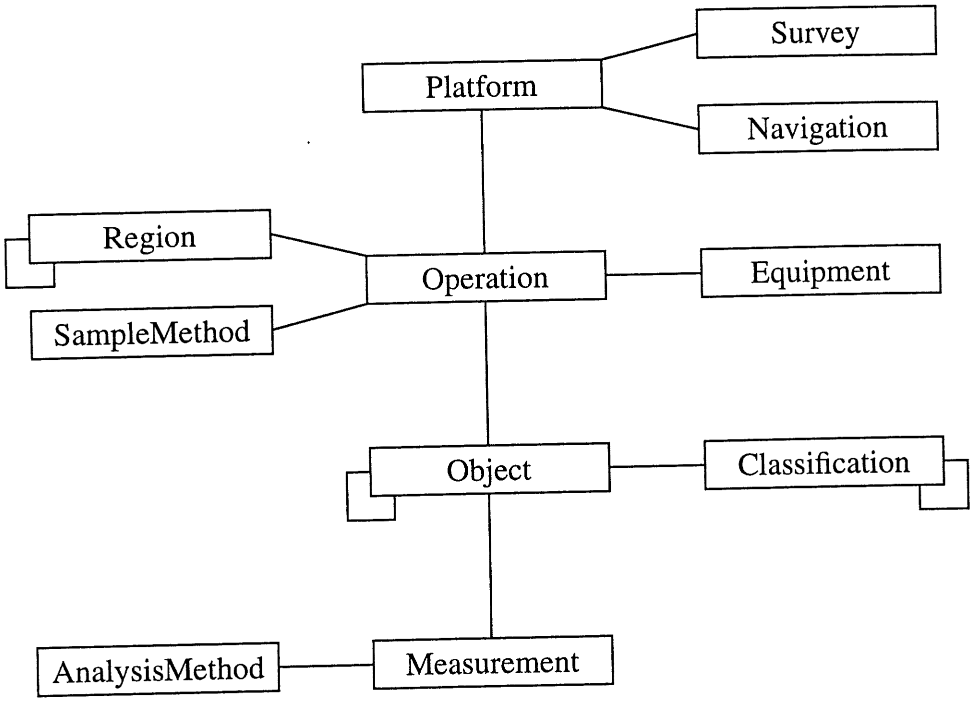
Tests of the database show that to attain a consistent and integrated database, it is necessary to establish general entity definition templates and detailed specifications of each data set based on these templates. As a consequence, all time-oriented data sets will be modelled according to the same basic structure ("platform-operation-object-measurement"). It is possible to have different representations of the entity definitions in the database without losing consistency, as long as a common basic key structure is used.

In order to codify the data according to different classification systems (e.g. taxonomy, regions and methods), special reference tables are included in the model. The reference tables are used to "look up" supplementary information about codified data, and are definitely an important part of the generic method of data representation.

### 3.1 The general data model

In the data modelling process, we discovered that most of the data sets had the same basic structure. The regularly sampled data all relate to a platform, to an operation, and classification and measurement of objects in the sea.

The main entities in the general data model can be illustrated in the following way:



**Platform** describes the unit from where the observations or measurements are made. Examples of platforms are research vessels, fishing boats, buoys and laboratories.

**Survey** holds information about the time, position, purpose and personnel concerning the investigation.

**Navigation** tells where a mobile platform is located (longitude, latitude) at a certain time.

**Operation** holds information on how, and under which conditions, a series of data is gathered. Examples of operational information are: starting time, equipment type used, state of the equipment, sample method used, regional localization.

**Equipment** is a reference list of various kinds of equipment used in data sampling, its capabilities, etc.

**Region** includes descriptions of naval regions and subregions. A region is defined as a series of latitude and longitude coordinates. Region is for some data sets the only locality information.

**Sample Method** is a reference list of methods used in the sampling of data.

**Object** is the observational target. There are two kinds of information about an object: classifications and measurements. A classification of an object is a categorization of the object according to a limited number of predefined classes or categories. A measurement of an object is an objectively measured value, in most cases an integer or a real value. In the general model, information on what kind of object that is observed is one of the attributes of the object entity, and any number of classifications or measurements on an object can be made.

**Classification** is a reference list of classifications or categories used to describe an object.

**Measurement** is an objectively measured value made on an object.

**Analysis Method** is a reference list of methods describing how to perform measurements on an object.

## 3.2 Modelling specific data sets

The general data model is designed to hold most types of scientific marine data. To accomplish this, it is necessary to define the entities in the general model, using general and less exact terms. The exact definition and placement of each data set in the model has to be specified.

The general model defines the operation as the act of gathering data starting at a certain time from a certain platform. In modelling the specific data sets, it is necessary to define what kinds of data and how many data series and data types that should be gathered in one operation. It is necessary to decide on what information about the operation is going to be stored.

The general model defines the object entity as the observational target, i.e. “anything in the sea that is going to be measured or classified”. The objects in each data set have to be defined and related. Some operations gather information about objects of the same type (a linear object structure), while other operations result in cross-related objects of many types (an object hierarchy). The relevant object structure has to be found and documented in each case.

The types of measurements and classifications that are relevant to each object have to be specified. The codification of measure types and classifications must also be designed.

Modelling a data set is done in four steps:
I Relating the data set to the general model.
II Applying a template structure: Find or define a template structure which the modelling of the data set will be based on.
III Specific modelling of the data set according to a template and the general model.
IV Placement in the database: The data can be placed in the generic or in the static part of the database depending on the nature of the data.

### I The general model

When inserting a new data set into the data base, it is checked if the data suits into the general model. If this is not the case for a major data set, the general model has to be revised.

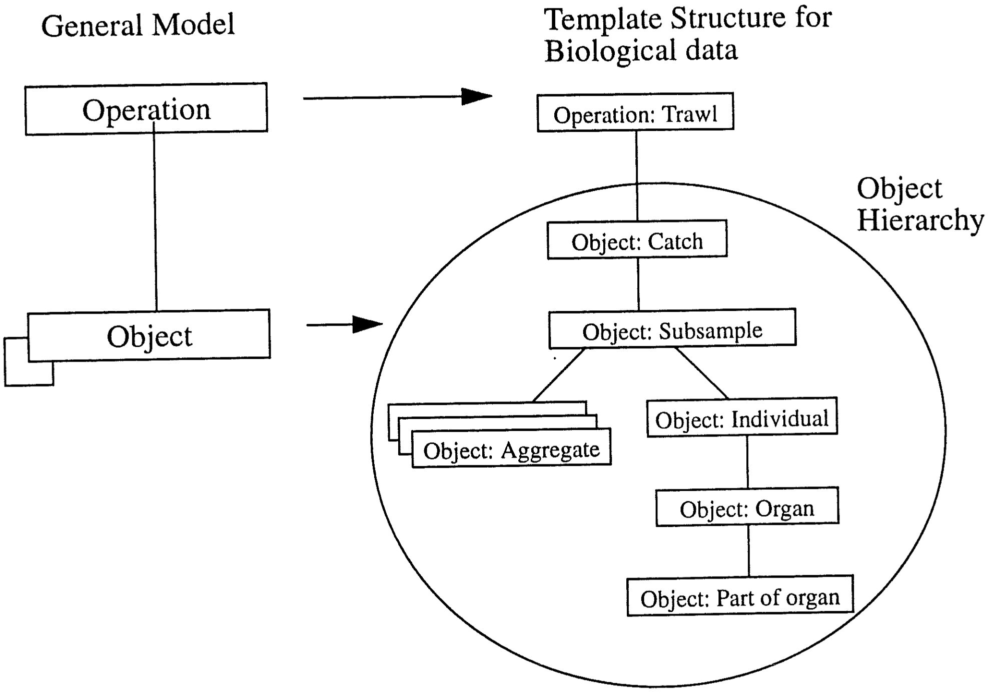
Most of the time series oriented data sets can be modelled according to the general model. Examples of data that currently does not fit easily into the general model are experimental data and observations made on the same object at different times (e.g. marking and recapture).

### II Template structure

Many data sets have more or less the same data structure. To achieve an integrated database, and to rationalize modelling of each data set, it is important to build data sets with similar structures based on the same template structure.

A few entity-definition templates are made to represent the different general data structures. In modelling a new data set we try to use an already existing entity-definition template. If none of the existing entity-templates suit, a new one is made - primarily with close resemblance to the other templates.

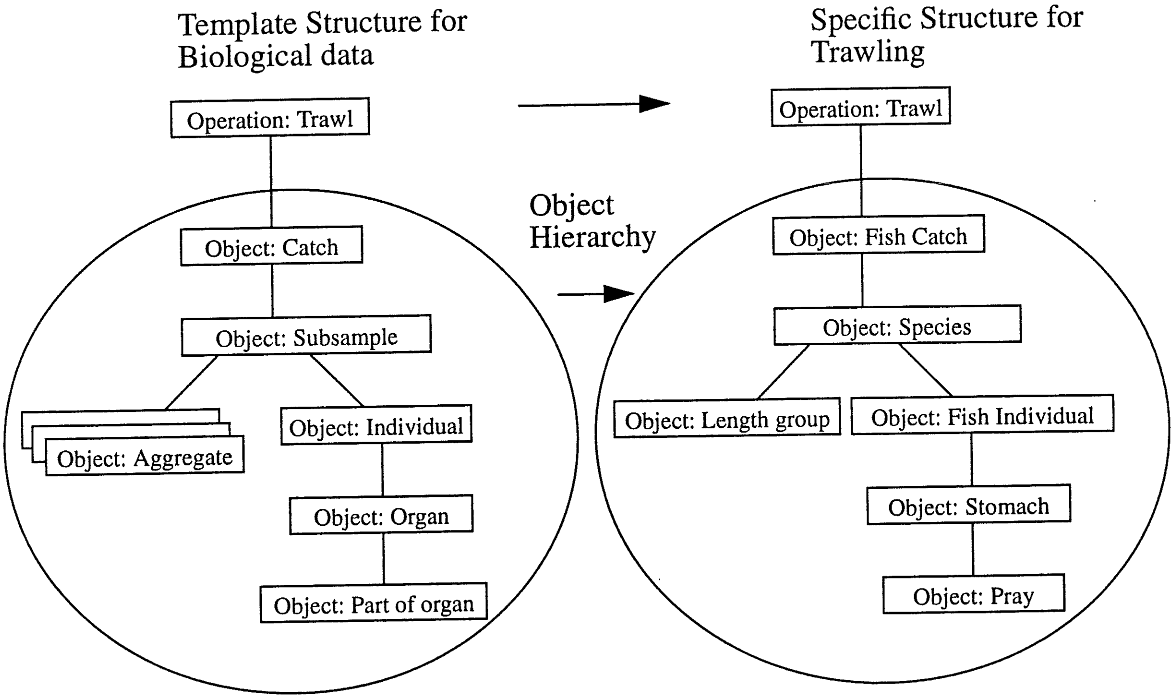
Relating the general model to a template structure for biological data:



## III Specific modelling

According to the entity-template, specific definitions of the entities in a data set are made. The definitions include a detailed specification of all the attributes of each entity and a specification of the relations between entities.

Relating the template structure for biological data to a specific structure for trawling:



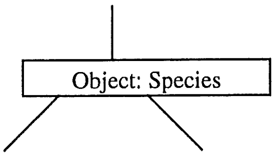
## IV Placement in the database

The entities are implemented either in the general or in the static part of the database. The data will be placed in already existing general tables if the entities are implemented in the general part. When localizing the data in the static part of the database, new specialized tables will be created. The data can be represented in both parts of the database, and there is a one-to-one relationship between an entity-implementation in the generic and in the static part.

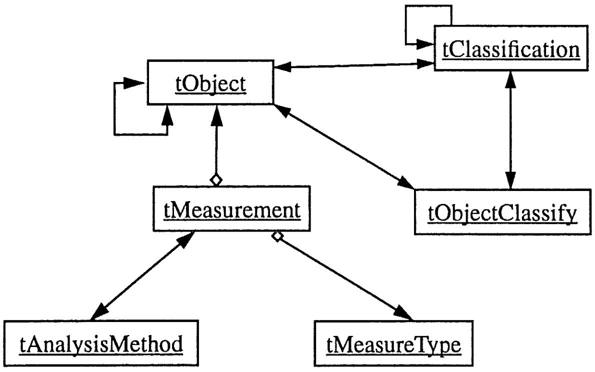
Objects with varying numbers of attributes can be placed in the generic part. Objects with a fixed number of closely related attributes can be placed in the static part of the database.

A generic realization of the Fish Catch object in the specific structure for fish trawling:

The fish catch object
in the specific structure for
fish trawling

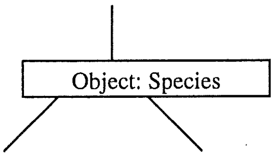


Ingres tables from the generic part of the
database holding fish catch information

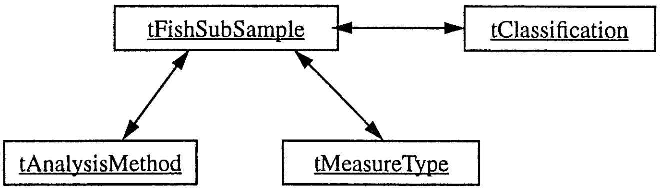


A static realization of the Fish Catch object in the specific structure for fish trawling:

The fish catch object
in the specific structure for
fish trawling



Ingres tables from the static part
of the database holding fish
catch information



# 4 Integration with the Bergen Echo Integrator (BEI)

The intention with the general structure of the IMR database is that it should be capable of holding any data used by the institute, or at least as many different data types as practically possible. At the start of the database project, there already existed a data model, the Bergen Echo Integrator Database, designed mainly for holding acoustic data. The main database and the BEI database are now coordinated so that the BEI database can be included as a subset of the main institute database.

The database is the core part of the BEI system. Volume backscattering data are logged from an echosounder via a Local Area Network (LAN) and are stored in high volume files on a hard disk at a rate of about 120 M bytes per day per frequency/transceiver combination. Each file set contains data from 5 nautical miles. Each of these file sets can be read into a colour screen picture called an echogram. Scientists scrutinize these echograms and add information using some of the features in BEI. The scrutinized data are stored with a suitable horizontal and vertical resolution into the BEI-database in depth channels for each species as area backscattering data.

The computers on board the research vessels will contain a copy of the main database. Most of the data sampled at sea, both those processed in the BEI system and other sampled data, will in the future be stored in the main database, and be available for further processing during the survey. On completion of the survey, this data will be transferred to the main database at the IMR. This will facilitate the use of identical database applications at the institute and on the research vessels.

# 5 Implementation

## 5.1 Realization in Ingres

The conceptual data model describes the data structures without taking the implementation into account. As a consequence, the model can be implemented in different database systems. The Ingres relational database system has previously been chosen as the data base management system at IMR, and the conceptual model was transformed into a relational model coded in Ingres SQL.

In the implementation process it was discovered that the set oriented relational database had some properties that were not considered in the navigation oriented conceptual model. In contrast with the conceptual model, in the Ingres relational database it is possible to access and relate data from all parts of the database directly.

## 5.2 Test of the database

In the last phase of the project a comprehensive test of the structure and efficiency of the database was accomplished. Various representative research data from the years 1988 and 1991 were converted, programs in ESQL/C were written and data was read into the database. To ensure that tests highlight complexity and time consumption as a user of the database will experience it, ten different search tasks were shaped by future users, in a range from "easy" (simple) to "heavy" (complicated), and programmed in SQL. Different data sets and structures were covered, and overall, the tests traversed most of the tables of the database.

With different setups and varying load the test programs were run singly, in parallel or assembled in combinations on one of the database servers, an HP 9000/755 computer. The performance was measured by the time consumption of the tests. The same tests were also run with various search criteria, alternative solutions, different amounts of data in the database and various internal data structures.

The test results show acceptable search time, varying from a few seconds to five minutes for tests run singly in sequence. Several searches run in parallel proved real time-sharing between running processes. Variation in logical and physical complexity between different approaches to the problems in the tests, resulted in significant differences in search time. Search time was also affected by the size of the database and type of data being retrieved and combined in the searches. In general, the solution method and realization of a search strongly affects the search time. As a conclusion, the search time is considered appropriate for most routine searches. Complicated, less frequent searches will be set up to run as batch jobs in the background. The extracted data from the searches will then be used in other application programs.

## 6 Further tasks

In the continuation of the work of getting the database operational, IMR in 1992 started four groups dealing with:

- Quality assurance of data
- Insertion of data into the database
- Database applications (data retrieval)
- Data modelling and database administration

The tests show a need for some modifications of the database, and some changes will be done to the data model. Also in the future regular revision will be necessary. This work will be taken care of by the data modelling group. At present work is in progress in the groups working on data quality assurance, inserting data into the database and adapting software applications to the database system. The database is planned gradually to come into regular use by the end of 1993.