ICES STATUTORY MEETING 1993 C.M. 1993/D:42 Sess.T

DEVELOPMENT OF A COMPUTERISED DATABASE SYSTEM FOR FISHERIES RESEARCH

D.R. Smyth

Department of Agriculture for Northern Ireland, Biometrics Division,
Newforge Lane, Belfast. BT9 5PX

INTRODUCTION

Data gathered from the various sources available for fisheries research represents a substantial investment of resources, in terms of both personnel and equipment. With the data being such a valuable commodity, it should therefore be utilised to the maximum possible level. In order to achieve this, an efficient and effective input, storage and retrieval system needs to be available. The existing system used by the Department of Agriculture was tied to highly structured data files that required programming knowledge to manipulate the data and was slow due the fact that only sequential access was possible. Any change in analysis or reporting requirements needed a thorough knowledge of the existing system and the appropriate level of programming skill, which could only be accomplished by trained personnel.

The decision was made to move to a relational database system, which would primarily allow users more control over the data by having easier access and also allow the possibility to integrate different data sets from the various sources where applicable. The system chosen was one already use by the Department called Oracle. This product consists of a central database engine that handles data storage and requests from the associated modules; these include a structured query language (SQL) module for interrogation of the database, a report writer, a data input package, a file input module and PRO*FORTRAN, a variant of the scientific programming language that can call database (SQL) commands. This combination offered the input, storage and retrieval components required by the system. While this may not necessarily be the best relational database management package available, it does offer great flexibility to the user.

The remainder of this paper will detail the data sets that were to be transferred to this system, the data input aspects of the work, how the system was utilised for retrieval and analysis of the data, and finally the future developments that are being considered.

DATA SETS

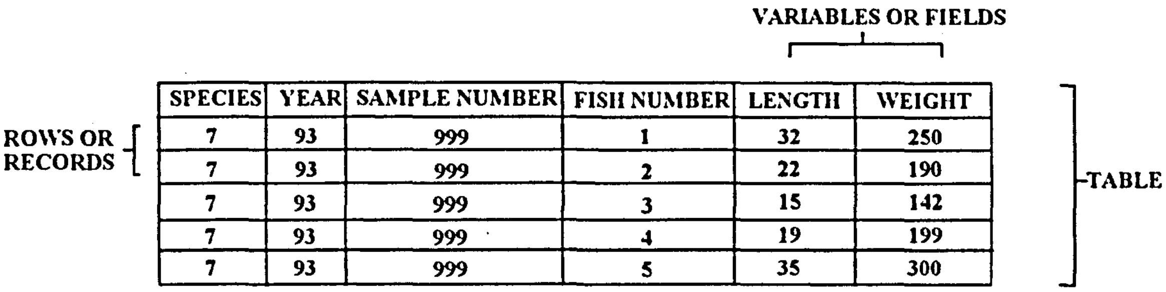
The main source of data was from the commercial port sampling work performed by personnel of the Aquatic Science Research Division(ASRD). Samples from the catches landed by the various vessels into the three main ports in Northern Ireland - namely Kilkeel, Ardglass and Portavogie - are purchased for the purpose of recording information on individual fish and the sample as a whole. This information consists of items such as species, date, main ICES area fished and vessel code and physical characteristics of individual fish such as age, length and weight. In addition, fish lengths are measured for the various species sampled during the landings into the ports, to produce data sets containing species, date etc., along with length-frequency values representing the number of fish from the sample recorded at the various lengths measured.

Similar data is also obtained from the fishery research surveys conducted by ASRD, however instead of the sample representing a subset of a particular commercial vessel's catch, it relates to the catch from a pre-defined fishing location, or station, in the Irish Sea, visited on these cruises. There is also considerable work being done on Nephrops (Nephrops norvegicus), with a system of data collection in operation for the vessels fishing for this species. This generates information on the date of fishing, area, gear, mesh size and length-frequency data for the various stages of maturity. The above comprise the major data sets currently being worked on, however there existed a substantial set of historical data for these areas, with information being held on the previous system from 1981 for the commercial port data and 1982 for the nephrops data.

DATABASE DESIGN AND DATA INPUT

The design of the database had to meet certain requirements. First, it needed to be easy to use, even to someone with relatively little experience using a database system. Second, it needed to be able to incorporate both current and historical data. Third, speed of access to the information had to be optimised to achieve the best balanced between storage requirements and retrieval times. Before proceedings, some explanation is required on the way data is held in the database. In Oracle, a number of variables describing a particular data set make a structure called a table, with each combination of values being stored in this table as one record. See figure 1 for a diagrammatic representation of this structure. These tables are entities in their own right, however they can be joined together if they possess one or more common, or key, fields.

The basic structure of the database was designed around two tables representing data from the port samples or cruises. One table, the sample table, contained information common to a sample such as species type, sample number, date and vessel code. The second table, the data table, contained the physical data, such as length, weight and age, particular to the individual fish in the sample. These tables can be joined by the year, sample number and species fields for the port data and species, year, cruise and station for the cruise data. Division of the data in this manner, called normalisation, helps reduce data redundancy by not duplicating common values and also improves data consistency by allowing the update of common information through the sample table rather than by the updating of the individual fish records in the data table. This is shown in figure 2. As can be seen in (i), the values species, year, sample, port and vessel are repeated for all fish, thus duplicating values common to all fish in the sample. In (ii) this common information is given once with only the key fields used for connecting the two tables duplicated. In relation to data consistency, to update say port code, each record who be required to be updated in (i), giving rise to the possibility of mistakes occurring, whereas in (ii), the information need only be update once.



SPECIES	YEAR	SAMPLE NUMBER	FISH NUMBER	LENGTH	WEIGHT
7	93	999	1	32	250
7	93	999	2	22	190
7	93	999	3	15	142
7	93	999	4	19	199
7	93	999	5	35	300

FIGURE 1. Table Structure.

The data were segregated in the database structure into historical (archive) and current sections for input reasons. Figure 3 shows the design that is currently in operation, with the boxes representing the tables holding the data. The four sets of tables in the current section are used to store data that is being input, updated or validated. The port samples, research surveys and length-frequency tables hold data on all the fish species being sampled, which at present include cod, whiting, herring haddock and hake. Once the current data has been thoroughly validated, it is transferred to the archive section, by a series of simple SQL commands, for permanent storage where no changes may take place.

(i)

SPECIES	YEAR	SAMPLE NUMBER	PORT	VESSEL	FISH NUMBER	LENGTH	WEIGHT
7	93	999	10	B1234	1	18	167
7	93	999	10	B1234	2	22	245
7	93	999	10	B1234	3	39	443

(ii)

SPECIES	YEAR	SAMPLE NUMBER	PORT	VESSEL
7	93	999	10	B1234

SPECIES	YEAR	SAMPLE NUMBER	FISH NUMBER	LENGTH	WEIGHT
7	93	999	1	18	167

FIGURE 2. Data redundancy and consistency.

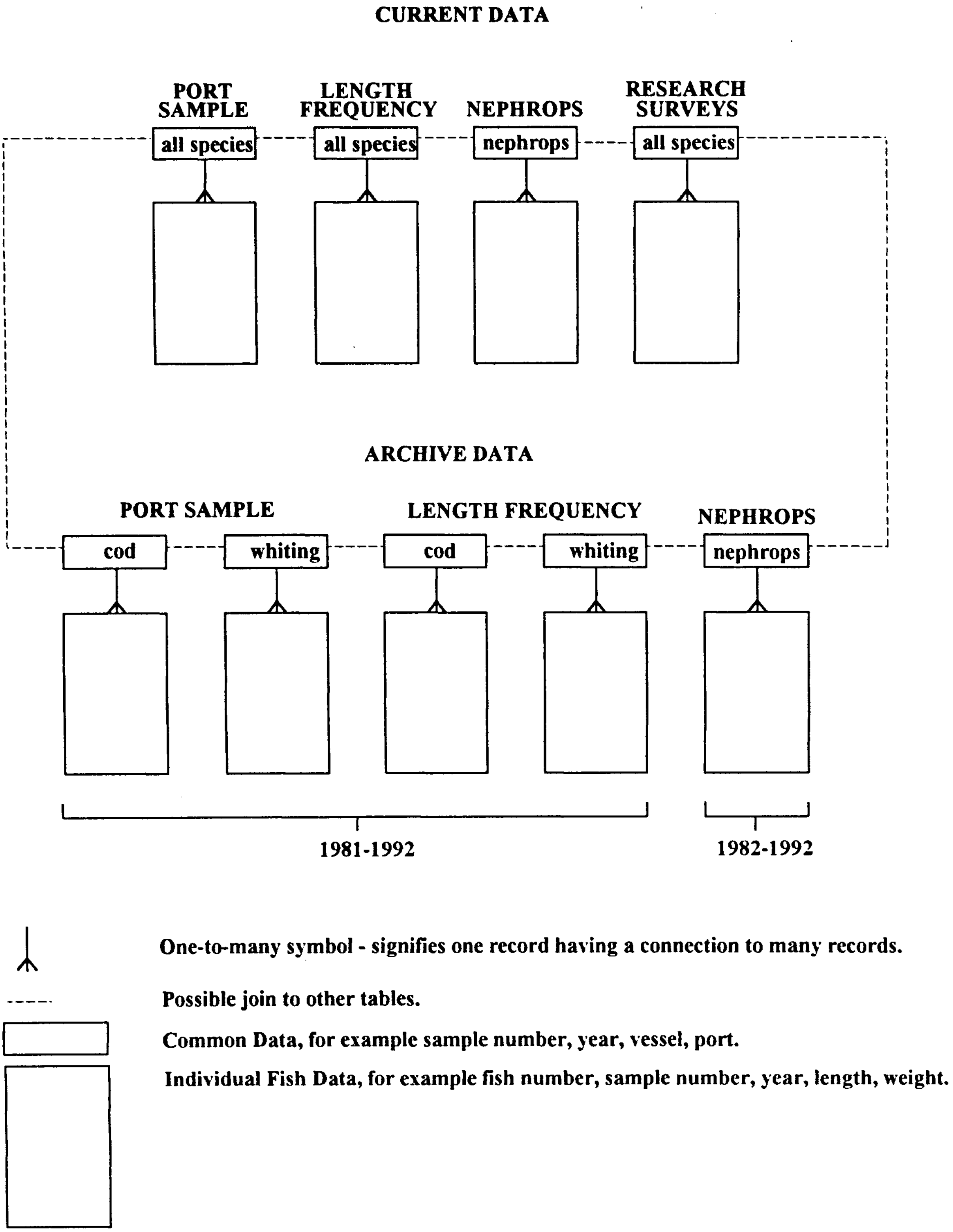
The archive section has the same basic structure as the current data, however instead of holding all species together, the data are split by species, for example cod and whiting, as shown in the figure. This was to increase access times for data extraction, as in most cases, only a single species was being investigated at any time. Storage was decreased slightly as a result due to the species field not being required in the archive data, although this was not the driving factor. With the relational design, all tables could still be joined together by criteria such as date, area, vessel or port if necessary so that for example possible species interaction data could be extracted. Only the research survey data have no archive section at present due to the recent development of this section, however one is being developed along similar lines to the port sample format.

Input of the data was also divided into current and historic sections. For the archive data, a large body of information was held in the form of formatted computer files that needed to be transferred to the database tables for that section. The programming module PRO*FORTRAN was used to read the archive data files and load the data into the appropriate tables in Oracle. This was achieved relatively easily by utilising the input sections of the FORTRAN analysis routines from the previous system, in combination with code developed in the SQL command language, accessible by PRO*FORTRAN. Validation of the archive data was performed in Oracle to ensure an accurate and complete data set using SQL commands to check specific values, ranges and calculated results. With the historical data successfully archived in Oracle, it immediately allowed a level of access and manipulation to the data previously not available to users.

The current data section was unlike the archive data in that data entry would be a continuous process throughout the year, rather than a one-off event. Full input, validation and reporting facilities were required to allow interactive storage and retrieval of the data, as samples became available. The previous system had offered an input system based on FORTRAN and the Datatrieve database, however this had proven very inflexible to any change in design and lacked a consistent user interface. Another problem was the fact that even though the data was originally input into Datatrieve during the data logging process, once completed it was stored in an ordinary ASCII data file for subsequent analysis, which defeated the purpose of using the database in the first instance. The solution, under the Oracle system, employed the SQLFORMS package to allow development of data input screens (see figure 4). Data may be input, updated or removed through the on-screen fields which correspond to fields in the current database tables with one or more input screens or forms for each of the tables used. Validation of the data is through devices called triggers; these are SQL commands embedded in a procedural language available to Oracle. They can be used to flag any discrepancies in the data entered by the user during an input session based on pre-defined limits, lists of values or calculations. The system offers extreme flexibility to the designer with changes being able to be made quickly and with relative ease. Functions available for the system include input, update and storage for all data sets (port, research surveys and nephrops data), interactive viewing of all sample information for both current and archived data and report facilities to give hard copy output of any sample. These are connected by a menuing system to enable easy navigation through the various options available.

DATA ANALYSIS METHODS

In terms of the database system, this aspect of the work is concerned with the possible methods for the retrieval and subsequent manipulation of the data. Analysis of the data stored ranges from simple report production to mathematical or statistical modelling, with the main effort directed to generating results for stock assessment at ICES meetings. There are four available procedures for data extraction, offering varying degrees of complexity, ease of use and expansion; these being SQLPLUS, SQLREP , SQLFORMS and PRO*FORTRAN. SQLPLUS is the easiest and quickest method of retrieval and analysis, by offering a simple command language that most people can become comfortable using. It has the facility to group data to give means, counts and variances for sets of data, or join data in various ways for simple report production. It has limited numerical abilities, that will cope with any basic calculations required. Most ad-hoc data requests are processed by this method, where applicable.



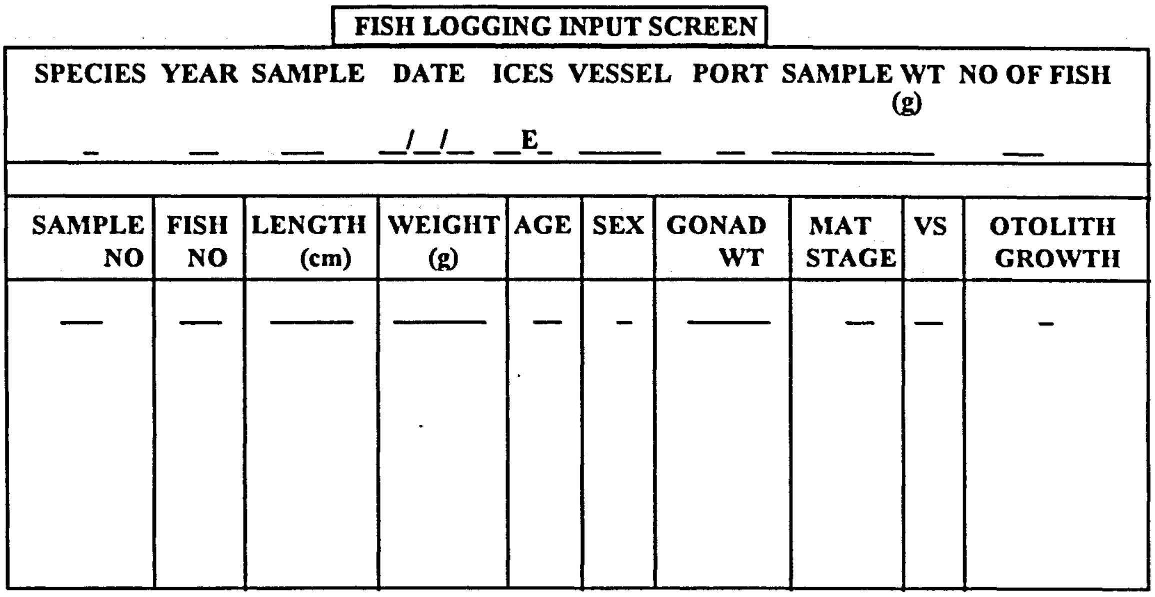
One-to-many symbol - signifies one record having a connection to many records.

Possible join to other tables.

Common Data, for example sample number, year, vessel, port.

Individual Fish Data, for example fish number, sample number, year, length, weight.

FIGURE 3. Database design for fisheries data.



FISH LOGGING INPUT SCREEN

SPECIES	YEAR	SAMPLE	DATE	ICES	VESSEL	PORT	SAMPLE WT (g)	NO OF FISH
_	__	___	__/__/__	__E_	_____	__	_________	__

SAMPLE NO	FISH NO	LENGTH (cm)	WEIGHT (g)	AGE	SEX	GONAD WT	MAT STAGE	VS	OTOLITH GROWTH
___	___	_____	______	__	_	_____	__	__	_

FIGURE 4. Example of input screen designed using SQLFORMS.

SQLREP is a dedicated reporting package that allows the user to control layouts to a high degree. It has slightly enhanced numerical functions for generating values such as running totals, but is still tied to the non-procedural SQL language. It requires some training to become familiar with the full functionality of the package, however once learnt, quite complex reports may be developed. The database systems makes use of some reports in its operation, mainly for data print-out purposes, although this is planned to be extended to incorporate more analysis procedures once firm specifications have been given. SQLFORMS, in addition to its main role as a data input device, can be used to perform calculations on the data held, displaying the information on screen, as well as offering basic data display. Where SQLFORMS differs from the previous two methods is in its ability to make use of the procedural language module, PL/SQL, available in Oracle. With this language, some very complex calculations and analysis may be performed, although a thorough knowledge of this product is required to gain the best usage of its facilities. Some preliminary work on catch per unit effort (cpue) calculations from the landings data have shown this approach to be flexible and powerful enough to compete with the FORTRAN system currently in operation. This has ensured the extension of SQLFORMS in its use as an analysis option.

The final and perhaps the best option for any type of data extraction and analysis is the PRO*FORTRAN package. A knowledge of FORTRAN is essential, along with the necessary procedures used by PRO*FORTRAN to access the database.

Together this produces a virtually unlimited array of analysis options, dependant only on the nature of the problem and the user's ability. This is the major analysis method used for the fisheries data. Analysis routines had been designed using FORTRAN for the previous system to generate age/length keys, catch at age data and other relevant information required by the fisheries scientists for their stock assessment work. These were able to be modified to utilise the database linkage available in PRO*FORTRAN, thus giving consistency in analysis procedures even with the change in data storage systems. This played a key role in the decision to accept Oracle, as by having the PRO*FORTRAN option, ensured that programs with many man-years of development invested in them did not have to undergo major revision to accommodate a new system. PRO*FORTRAN indeed enhances the development of new analysis routines and extensions to the originals as the programmer does not have to be concerned about how the data is stored, only where it is stored and under what field name. Currently work is being considered in connection with major enhancements of the information retrieval for stock assessment work, by bringing in data from other areas; something that would have been practically impossible under the old system.

FUTURE DEVELOPMENTS AND DISCUSSION

Much development is still possible by extending the scope of the database to encompass the other data sets available for inclusion. The main data set that is being considered is the landings information recorded at the registered port throughout the UK. It contains all manner of data connected with a vessel's fishing trip, including items such as length of time at sea, area fished, vessel details and catch size, cost, freshness and quality. These data have been under-utilised for a considerable length of time, mainly due to the size of each year's data set (approximately 20 Mb) and the lack of available development time. Design of the database structure for this information is already underway, with a number of possible designs currently being reviewed. Once the system is in operation, it will make the data more amenable to access and therefore allow easier development of analysis programs.

Advances in technology have opened the possibility of a management information system (MIS) to enable personnel with some SQL knowledge the ability to interrogate the database at a reasonably sophisticated level. Although no firm work has been initiated, the general idea would be to have some form of window system resident on a PC, such as a database development package, with a network connection to the central database storage area. The user could interactively select various fields from the database, by different criteria, in order to perform report production or analysis. This would allow access to the data by personnel of any level of expertise.

The efficient and effective management of data is of prime importance to any organisation that is required to handle and process information from the many and varied sources. The implementation of a relational database design model to fisheries data has proven of great benefit in terms of the access that is now capable and the level of control over the data offered to the user. Development of the system meant a large investment of time and resources, however this has already started to pay off due the reduction in time spent on programming new analysis routines.

ACKNOWLEDGEMENTS : Thanks are extended to Mr. Paul Toland of Biometrics Division for his programming assistance throughout the development of this system and to members of the Fisheries Section of ASRD for their helpful suggestions during all stages of the work.