The research potential of shared information resources

B. W. WICKHAM

New Zealand Dairy Board, Hamilton

ABSTRACT

Recent developments in the technology for storing, retrieving and communicating data have lead to the possibility of greatly expanded data sharing. The advantages of extended data sharing in the animal based industries of New Zealand include the possibility of conducting research studies previously considered not feasible because of the resources required to collect and collate the needed data.

With the sharing of information between farmers, Government agencies, research institutions, servicing organisations and advisory services, it is possible to forsee research studies of considerable benefit to the primary industries in New Zealand. This paper will examine the concept of shared information and provide illustrations from the New Zealand industry of the research potential.

INTRODUCTION

Research typically involves the development of models to explain various phenomena. These models are arrived at and enhanced by a careful examination of the way they fit observed data. Under the classical experimental approach, the scientist controls the testing of his models by designing experiments, collecting the appropriate data, and examining its 'fit'. For the livestock industries of New Zealand, this approach is very expensive, because animal and land resources have a high commercial value. At the same time, this commercial value has given rise to the collection of large amounts of data for management purposes. Much of this data has potential value for research.

SOURCES OF DATA

Large amounts of data are routinely collected as part of normal commercial activities in New Zealand's agricultural industry. Data collection occurs on individual farms, where, for example, records are kept of:

- parentage of individual animals
- calving details
- premating oestrous behaviour
- mating details
- pregnancy status
- mastitis treatment
- drying off
- paddock grazing
- fertiliser application
- bloat prevention
- spore counts
- receipts and payments

These examples illustrate the broad range of information as well as the existence of detail which could be of value in animal production research. However, to this must be added information collected by service organisations. For example, Livestock Improvement Associations (LIA) which provide services to over 75% of dairy farms, collect information relating to:

- herd testing and artificial breeding services
- individual cow production of milk, fat, protein and somatic cell concentration
- artificial insemination of individual cows
- animal parentage
- calving

The data collected by the LIA are already used extensively in research studies but there are substantial areas of research which would benefit from access to these data. Other organisations which provide services to dairy farms and collect data include, breed associations, veterinary clubs, dairy companies, stock and station companies and Government departments.

The Dairy Board's Farm Production Division, in addition to collecting data from farms on cows in its sire proving scheme, is responsible for summarising data collected by the LIA and providing statistics which are used in industry decision-making. An example is the genetic evaluation of bulls and cows. These statistics are valuable in other research studies. The main point to be emphasised is that there is a very large amount of data being collected by a wide range of organisations which has substantial value for research purposes even though it has been collected primarily for other reasons.

AVAILABILITY OF DATA

There are 5 main factors which influence the availability of these data for research purposes.

1. Researcher's Knowledge

Research workers are not always aware that data exist which may be of value in their research projects. There is little published information available which describes the detail of data collected by organisations. The relevant data will not be sought if the researchers do not know of its existence.

2. Diversity of Origin

Data useful for research purposes may be spread over many different organisations. This presents an obstacle in that each organisation must be approached and permission to access the data sought. Delays normally result and it may be possible for organisations to effectively veto the research project by refusing access.

In the dairy industry there are 17,000 farms, 6 LIA, 7 breed associations and the Board's Farm Production Division. Depending on the type of data being sought, it may be necessary to deal with 1 or all of these. From a research access point of view, a single organisation providing access to data would be highly desirable. However, it would still be necessary to ensure proper authority for the release of the data had been given.

3. Policy of the Organisation

Some organisations refuse to make data available for research purposes. Although this is extremely rare in New Zealand, issues of confidentiality and commercial sensitivity may prevent data being made available. To overcome these difficulties it is essential that the organisation conducting research be able to provide adequate assurances to the providers of data.

An illustration is provided by the way in which the dairy industry is able to obtain access to herd test records. A condition of the herd testing service is that a copy of the records be provided to the Dairy Board for use in research. In this way, the Board has been able to conduct a very large number of research studies by accessing data collected as part of a normal commercial service.

4. Method of Storage

Data storage is in the middle of a major technological revolution. Storage in the past has been based largely on written records on paper or microfilm. These media make data retrieval time consuming and costly. This is particularly the case in research studies where the type of detail required means accessing the data in a sequence different from that in which it has been stored. Typically, data extraction has been labour intensive and often requiring highly skilled staff who have a very good working knowledge of the stored data. In combination these factors have resulted in data extraction being very expensive.

With the introduction of computers, there has been a substantial improvement. Since the data are held on machine readable media they can be accessed rapidly and at relatively low cost. However, the organisation of data on computer files can be such that a major programming effort may still be required before the required data can be extracted.

5. Quality of Data

Once retrieved, the data may be unusable for research because of quality factors. For example missing sire codes may be so frequent that a genetic study is ruled out. Data quality is closely related to the emphasis placed on quality during the initial data capture. In this respect, there may be a conflict between the quality required for research and the associated costs incurred at the terms of data capture.

TECHNOLOGICAL DEVELOPMENTS

Recent developments in computer technology are resulting in dramatically reduced costs for data collection, storage and retrieval. These developments in conjunction with new approaches to organising the stored data hold the promise of cheaper and faster access for research studies.

As can be seen from the preceding discussion, the organisation of stored data often presents an obstacle to accessing the data in ways which the designers of the storage method had not anticipated. This problem is being approached in 2 ways. Firstly, increasing emphasis is being placed on the development of logical data models and secondly, physical file designs are moving towards the use of relational databases.

Logical data models provide a statement of the meaning attached to data items and the way data items are related to each other. These models are built largely without reference to their physical implementation. Emphasis in building a data model is placed on identifying the data needed to achieve an organisation's objectives. In the case of the dairy industry, this exercise involved examining the objectives of all organisations sharing in the development. That is, dairy farmers, LIA, breed association and the Board's Farm Production Division. The resulting data model contains approximately 1,500 data items grouped into 300 entities. This model consists of a dictionary of data item definitions and a map showing the relationship between groups of data items (entities). The value of a logical data model is that its development requires a thorough examination of all aspects of an organisation and the data needed by each section for its operation. As a result of the development of a data model, it has become very clear in the case of dairy industry that many of the data needed originate from dairy farms and are then shared by other organisations.

This sharing logically extends beyond the organisations involved in the development so far. From a research point of view, the data model allows data items of potential value to be readily identified. It also provides a formal method for introducing new data items which may, for example, arise as part of a research study. Typically, the consequences of introducing new items is minimal. This is because the model reflects the logical organisation of data. Data models make few assumptions about the way data are to be accessed and thus leave open all possibilities involving the data they contain. In this way it is conceptually possible

to obtain access to any data items in any way. However, data models are conceptual only and may not be able to be physically implemented with current technology.

Relational database theory provides a theoretical basis for allowing logical data models to be physically implemented. However, there are few working examples in existence and today's computer systems are still being built as close approximations.

In the case of the dairy industry, a database is now being built to closely resemble the logical data model. This building process involves selecting parts of the model to be implemented, choosing amongst available technologies, designing files, designing data maintenance procedures, designing reports and writing programs. Where the database involves several organisations, it is necessary to establish clearly the responsibility of each in terms of their ability to add, read, modify and delete data. Once this has been agreed it is relatively straight forward for several organisations to share responsibility for data held even on the same physical files.

Extensive use is being made of database management software, which allows programs to be written independent of file design. Programs are being written in high level languages which allow development to proceed much faster than with languages such as COBOL. The resulting database is being shared across several organisations. Over the next 12 months, it is planned that the development will replace virtually all existing computer systems. Thus by the end of 1984, the combined organisations will be in a position to provide greatly improved access to existing data.

The development of data models and building of databases provide a degree of centralisation of data storage and system design. This move is being countered by the very rapid development of personal computers adapted for farm use. Software is being developed by many individuals and organisations with a view to making sales to New Zealand's farmers. This software is being designed for use on particular machines and to provide certain data recording and reporting functions for individual farmers.

There is a need for the data collected by farmers to be made available to central databases and vice versa. These needs arise out of the benefit the farmer receives from having access to comparative information and the benefit he receives from research carried out using his and other records. Development in the sharing of data between databases and farm microcomputers is in its infancy but it has the potential of greatly extending the range of data available for research purposes.

Similarly, the sharing of data between organisations servicing agriculture is in a very early stage of development. In addition to research benefits a more extensive sharing of data is likely to provide opportunities for improved services and cost savings.

RESEARCH POTENTIAL

Illustrations of the research potential which can result from access to shared information can be found in the dairy industry. These include:

1. Studies of the genetic basis of production, reproduction, conformation, physical traits and disease susceptibility in dairy cattle.

This example is an extension of work already carried out by the Board's Farm Production Division. The extension into animal health aspects is based upon the assumption that more extensive recording of disease diagnosis, testing and treatment could be readily achieved. The Board's databases could provide storage of these details collected from both farmers and veterinarians.

2. Studies of other factors which affect production, reproduction, conformation, physical tests and disease susceptibility of dairy cattle.

This example is similar to (1), but illustrates another important aspect of the database, that it will provide a simple mechanism for identifying a wide range of environmental factors such as type of shed, geographical location of farm (by map reference), start of mating, origin of cows and so on. As further organisations become involved, it is easy to envisage data on milking machine characteristics, milking times, soil fertility and fertiliser usage becoming available.

3. Studies of the relationship between pasture management techniques and productivity.

As farm management data become increasingly computer based, it will be possible for pasture management data to be collated from many dairy farms. In conjunction with other data, they could be used as the basis for studies of the consequences of alternative management strategies. In this case, a shared database provides access to information which would be useful in estimating the spill-over effects of various decisions.

4. Mixed experimental/commercial collection of data for research studies.

In a shared information environment, it will be possible for experiments to be designed and conducted on commercial dairy farms, with a large percentage of the data collection being handled by 'normal' non research means. The data specifically associated with a particular experiment could be recorded by expanding the database appropriately. In this way, large scale experimentation can be conducted with relatively little added cost for data collection.

CONCLUSIONS

Research covers a very broad range of investigatory work, but regardless of its level of sophistication or the nature of the question being examined, there is always a need for data. Developments in the field of information processing promise greatly improved access to data for research studies. In the New Zealand Dairy Industry, the first results of these developments

will occur during 1984 and 1985. The benefit of the resulting research is likely to extend beyond the farmers and organisation providing the basic data.

Data sharing between organisations is in a relatively undeveloped state and there is a clear need for investigation into its feasibility and long term benefits. This applies particularly to the sharing of data between farmers and central organisations.