BRITISH NUTRIENT DATABASE

Derek D Singer, Ministry of Agric., Fisheries & Food, UK

BACKGROUND

The UK Nutrient Data Bank started life in the 1920s with the work of Dr. McCance and Widdowson at King's College, London. Their work was funded for many years by the Medical Research Council (MRC) and published in scientific journals and as monographs. About 1940, the first complete set of data was published in book form but after 1960 there was little further development for a number of years until the Ministry of Agriculture, Fisheries and Food (MAFF) required updated information for its own purposes. Dr David Southgate at the Dunn Laboratory ,Cambridge, was commissioned by the Ministry with the agreement of the MRC to spend some time on new tables. The analyses were funded by MAFF and carried out mostly at the Department of the Government Chemist, where incidentally, at that time I was employed. More recently, work on the tables has been completely taken over by MAFF and is the concern of my colleague Dr David Buss.

Since 1976, I have been involved in the 'computerisation' of the data, particularly for use in the Food Science Division of the Ministry where it is applied to the many dietary surveys sponsored by the Department.

Prior to my involvement in the data bank, the data had been published on paper tape at the cost of about $250 in current values. Unfortunately the format was complex and a number of errors were incorporated. All copies have now been sold.

Our original programming was carried out on a small WANG 2200T machine and then on a WANG MVP using WANG BASIC2. The facilities included the ability to select foods by 'code' or name, to store recipes and diets and to use the data bank directly in the processing of surveys. A rapid string search of food names was available to help the user select the appropriate food without recourse to a written list.

Since 1984, the programs have been completely rewritten in FORTRAN 77 for use on a PRIME computer.

In the meantime the UK Government has been reviewing the cost effectiveness of many departmental activities and also the dissemination and possible profitability of Government information. One outcome has been that departments are now being exhorted to review ways of profiting from the sale of information where that information is a 'negotiable entity'. For

example, the contents of our annual 'National Food Survey' can no longer be quoted in answer to 'phoned enquiries; the advice given is that 'the answer to your query will be found in the 'National Food Survey' price $2.50 published by Her Majesties Stationery Office"!

However the situation in 1983 was that many unauthorised copies of the data existed, some corrupted by insertion of non-official, unapproved figures. This data was often sold or freely diseminated via the invisible academic network, 'warts and all', without authorisation, together with computer programs of sometimes dubious value. Our publishing department had no immediate plans to publish the data in machine readable form in spite of the many requests received. Meanwhile sales of the book were healthy in the UK and in Europe, where it is handled by Elsevier of Holland (who market it at a price approaching three times that of the UK price).

Much of the data, although far from useless, was out of date. Although new analyses had been carried out and were used in our own work, the incorporation of the new data into published tables was becoming increasingly difficult as staff were heavily engaged on other activities. Accordingly, as Head of Information and Computing Services within Food Science Division I began to investigate the possiblity of contracting out not only the publication but also some part of the compilation of the tables.

For a time I had no success in finding a commercial publisher with the necessary all-round expertise but in 1984 I was asked to join the Users' Advisory Committee of the UK Royal Society of Chemistry. The Society is one of the largest publishers of chemical information in the world (it is said second only to the American Chemical Society) and earns the highest regard of the scientific community. It is experienced in the construction of databanks and databases as well as in the publication of hard-copy. In common with Government Departments and most UK academic and professional institutions, the financial climate was at that time was inducing it to take a firm commercial view of its activities as far as it could in the light of its status as a learned body.

At my first attendence at the Advisory Committee a request for suggestions for new databases or databanks was made and as an outcome my Division commissioned the RSC to carry out a market study of a new nutrient databank in the UK with some reference to a European Data Bank. The study showed that a new databank is a viable commercially and as a result the RSC will supply a nutritionist to work within our own Nutrition Branch in selecting and validating the data and will construct the databank in liaison with my own Branch for publication in various forms.

The discussions between the RSC and MAFF have been spread over 3 years. During this period we have talked with INFOODS and

EUROFOODS and attended their meetings. What follows is a selection of the problems and questions that have arisen from these forums in as far as they affect the UK scene together with a summary of the new features of the tables.

USERS & USES

Nutrient databanks are unusual in many respects.

First the data is used by remarkably disparate sets of users. Users are found in Government, academic institutions, industry, medical institutions, and within the general public and those who serve the general public.

Respectively, these organisations include departments within national Government (eg Departments, Armed Forces, prisons) local Government and the International Government (EEC and UN): schools and universities: food processors and distributors: doctors, dentists, epidemiologists and dieticians: writers, broadcastors, journalists, pressure groups, advertisers and the general public. Of late a new area has become important - the law - since nutritional labelling is now under active discussion.

These users have differing needs arising from their differing activities. In order to decide how to meet these needs we considered aspects of Information (Publishing), Computing and Nutrition and the interactions each have with the others.

INFORMATION/PUBLISHING PROBLEMS

A book is required. Should this book be bound or loose leaf? Loose leaf is convenient if updates are to be issued at comparatively short intervals - which has not been the practice in the past - but the cost is higher and the pages more subject to damage or loss.

What should the cost be? Clearly the publisher must make a profit - or at least not a loss! But should the cost reflect the full cost of the all the work contributing to the publication, including the analyses? Undoubtedly were we starting from nothing the costs of the analyses would be many millions of dollars. In the UK during the 70s the cost of analysing about 200 foods approached one million dollars at 70s monetary values. We conclude that the cost of analyses cannot influence the price of the publication.

In the UK the most highly regarded and trusted data is produced by government for its own use and other users have no direct influence on distribution, costs and copyright. A number of

academics have taken the view that the data should be freely available. These persons would take a different view if they were asked to work for nothing, or to write books for free distribution. In the light of the high costs of analyses it seems reasonable that users should at least bear the costs of production and distribution.

Without control by the author department and the publisher the data would become diluted and corrupted by 'foreign' information - an obviously undesirable happening. The user will be protected from spurious data by the RSC copyright and licensing policy.

Who owns the data? Although we are now transferring responsibility to a commercial body, albeit one who is a learned society, all work sponsored by the UK Government is the copyright of the Crown. The new data will be licensed solely to the RSC who will be empowered to sub-license it as occasion demands. The data will be available internationally wherever there is a market and should be easily available via normal commercial outlets. It is likely that software houses will be granted sub-licenses to sell the data with their programs.

What should the data contain? The Ministry wishes to expand the data-bank to embrace the hitherto unincorporated data in its possession. One of the valuble properties of M&W is that the greater part of it arises from original analyses and this feature accounts for the supreme trust placed on it by users. Much other data is available but if extraneous data is included it should be tagged as such and it is likely that the data originating from the Ministry will be published either separately or as a clearly distinguishable part of the complete set.

The question of sets or subsets are important. The new databank could be very much larger than the old and its price accordingly higher. Not all users will require the full data. It is therefore likely that subsets will be available containing less detailed data or restricted to certain classes of foods in order to satisfy as many users as possible and take advantage of the full market. Publications might range from many volumes for research workers to pamphlets or booklets for the consumer.

Data tagging could be extensive. The number of samples is already a feature, but should the range of results be given? Should details of the samples be provided? This is discussed under nutritional requirements but involves agreement with, for instance, food processors if brand names are to be included and with authors and other publishers if literature is quoted. Data tagging might include the analytical methodology and relevant literature references.

Such data tagging adds considerably to the costs of the databank. It is likely that this information will be stored by the RSC and ourselves but only available on special request on

computer readable media and of course at an appropriate price. In general, most users do not require detailed or extensive data; but a few might want everything available. These latter must expect to pay for the extra service.

COMPUTING PROBLEMS

There is an obvious market for the data in machine readable form. However the RSC market survey indicated that little market existed for an on-line service. It is possible that this result stemmed partly from a current shortage of funding for use of such services by the users within the UK and ignorance of such services by many questioned. However no firm decision has been made.

Users want the data in machine readable form for use in-house. Machines may be main-frame, mini- or micro, made by any of a number of manufacturers and using different media and formats. The data format could conform with one of the international standards, ISO or EUSIDIC.
The RSC will probably make tapes available in the format used by the large data base hosts and for mainframe and mini-computers such as the IBM, DEC, PRIME, HARRIS and for IBM, ATARI and other popular micros.

Purchasers will need to sign the usual agreement to protect the vendor against unauthorised copying. Copies for use on micros could contain internal protection.

It is unlikely that the RSC or the Ministry will provide programs to utilise the data and agreements with other organisations to provide various facilities for sale with the data under license are under discussion.

With some reluctance I must deal with coding. The need for a food code arises partly to enable data to be stored and programs to be written on small computers, partly for easy data capture during surveys, partly for easy programming to process surveys and partly for use as a mnemonic by everyday users.

The problems are manifold. Nutritionists like food groups. Most foods fall clearly into groups - cereals, milk and milk products etc. But is baby milk a health food or a milk product? Is a table jelly a dessert or a meat product? Anomalies and semantic considerations pose enormous difficulties and intellectual problems.

Do we use a hierarchical system, a network, a thesaurus or a faceted system? The problem is one of designing a system for all users and all uses. The more sophisticated it is, the less use it is as a mnemonic, the more complex it is for the non-professional computer programmer, and the higher the redundancy - which has to be paid for. The less sophisticated it

is, the more anomalies and ambiguities will arise. The code that most suits data capture and entering survey data (data preparation) might not be optimal for data retrieval and processing. Experience with processing many UK Government surveys covering nutrients, contaminants and additives leads me to believe that it is best to design any coding uniquely for the particular survey. The results are better and less time and money is expended.

The USDA is to be congratulated in evolving its faceted system but the effort required to code a large data bank will be high and sooner or later someone will want to access and process the data in some way not covered by the system. I am still thinking how to apply it to the two hundred or more French cheeses or the many varieties of UK biscuits (cookies). Such systems possess high redundancy which can not be afforded by the average user. For use by the source organisation it is however not without attraction and resembles a system we had been considering for some years which uses multi-valued and associated fields and was intended to be universally applicable to an integrated nutrient/additive/contaminant data bank.

The new published version of the UK data bank is therefore likely to use sequential numbering, perhaps broken to signify the broad food groups. In other words there will be no coding system which contains information of high significance. However programs could employ inverted files containing descriptors appropriate to, let us say, the components of composite foods. A cheese file might be include the key word pizza and an associated field provide the cheese content. Searchers would then be able simply to ask for all foods containing cheese and to carry out calculations on cheese consumption with the certainty that pizza would be taken into account alongside other similarly coded cheese-containing foods. Inverted file software for micro-computers has become readily available over the past few years and the widespread use of Winchester disks offers the facility to store large inverted files. Recipes may conveniently be handled this way. Users can easily construct their own inverted files, or concordance to use the information terminology, to suit their own purpose. Such a system could be thesaurally based, but I do not support, in this application, a highly structured system.

NUTRITION

The data contained within a nutrient data bank is not immutable. New food products are developed such as snack foods, confectionery and desserts. Existing foods change in composition as different varieties of plants and animal are farmed. Animals are butchered in different ways. Fortification practices alter and composition may change with processing and packaging techniques. Manufacturers change constituents as the price and availability of alternatives fluctuate. Moreover, even in the small geographical area of the UK, food processors may change

the composition of their products according to regional preferences.

As with all sciences, the emphases of nutrition change with time. Different data are required and some data decline in importance. Currently, except for some clinical purposes, there is less interest than formerly in amino-acids but more in sugars and individual fatty acids.

Compounding these problems, new analytical methods arise and occasionally new substances evoke interest.

Our own nutritionists are well placed to take note of these changes and, where relevant, appropriate values are used in official surveys. Table users outside of Government are generally not so well placed.

Since 1978, when MAFF assumed the sole responsibility for the tables, a large number of foods have been analysed. These analyses form part of a rolling program centered on the major staple food items, foods with high concentrations of selected nutrients and foods that are important in the diet of selected population groups, such as immigrants.

Supplementing this program are analyses of selected foods for specific nutrients in more detail. As a consequence we possess extensive data not presently incorporated in published tables. The data refer to more nutrients, some in greater detail and in some cases to regional and seasonal differences.

The additional nutrient analyses will fill many gaps in the tables and the additional nutrients selenium and iodine have been analysed in some foods.

The greater detail includes carbohydrates and Vitamin A. The current published tables cover available carbohydrate, sugars (lactose when appropriate), starch and dietary fibre. For many foods we now have data concerning individual sugars (sucrose, glucose, fructose, lactose, maltose etc). The dietary fibre figures are supplemented by components namely non-celluosic polysaccharides (hexoses, pentoses, uronic acids) cellulose and lignin. Similarly Vitamin A is covered in the current publication by retinol and carotene whereas we now have data on trans-retinol, 13-cis retinol, dehydro retinol, retinaldehyde and B-carotene. Regional and seasonal figures for the composition of bread and milk in seven different regions of the UK are now available and for potatoes we now possess new data for different varieties, regions and seasons. Some of those data has been published in scientific articles and some not. None appear in the published tables.

We will also supply the Society with extensive data on lipid contents of food not hithertoo generally available in easily usable tabulated form.

As already implied, it is hoped that manufacturers will also supply their own data for incorporation into the new tables. Many manufactures, perhaps most, are very willing to make data available.

Nevertheless there still exists many gaps. We ourselves draw on several sources including other analyses carried out in the UK, general literature values and the food composition tables of other countries - although data from the latter sources need careful consideration of their relevance.

We have briefly considered inclusion of bio-availablity. At present we think that the data are too fuzzy and open to mis-interpretation to include in the tables.

In constructing tables the MAFF has borne in mind the many uses and users both outside and within Government. There are clearly problems in constructing a table to suit all needs. The increasing interest and endeavour in nutrition and the generation of more foods and data, lead to many different but equally valid figures becoming available. We as the producer of the tables are in a position to select values according to application in the light of our background knowledge. However other users are not in this position. The current tables provide a 'representative' value. If we were to provide all the values in our possession how would the users select which to use, particularly if they did not have information either on the data or on the sample with which to base their selection? One does not buy powdered milk labelled 'from a Friesian cow, Southern England, April 1985' although in some cases such data would be very useful. In our survey of the diet on the Orkney Islands we used data on Orkadian foods. What use should be made of this data in the tables? A good deal of instinct and value judgment together with knowledge of market shares and the relevance of data spread is used in the choice of the representative values but users should be aware of uncertainties.

Inevitably, a nutrient data bank is more than a mere list. Numerical data are complemented by text which is important for users to read and understand. As the data grow in extent the significance of the qualifying and modifying information also grows. Whilst some of this textual information is amenable to tabulation or tagging eg analytical methods, preparation etc. some is not. Users must beware of using computerised tables without due observation of the context in which the data is derived and the text contained in the printed versions.

CONCLUSION

The new UK data bank aims to provide exended data, more rapidly updated than hitherto, in a variety of forms and formats and on a variety of media. Inevitably the new publication will be more

costly than the existing one but subsets and condensed versions will be available. The Royal Society of Chemistry will be able to consider special requirements; it will set up a small advisory group and probably a User Group. The Ministry will continue to play the leading role in the provision and validation of the data and will ensure that no loss in quality will result from the commercial approach.

Finally I must mention the interaction with the international scene. The RSC market study revealed little interest in composite international tables. The UK is no longer the 'tight little island' it once was. The significant immigrant population and membership of the European Common Market has led to importation of many new foods, and people now travel more widely than they once did. This lack of interest does not therefore stem from insularity. Rather it is a indication that foreign food tables have insufficient relevance to the situation in the home country to justify any merging. When a demand arises it is not difficult to obtain a value - and for the large majority of users demand arises very infrequently. Interest is shown only by a few research workers, epidemiologists and those in other very specialised specialised areas. The difficulties of table production on a national scale are multiplied enormously on the international one. Whilst I admire the efforts of INFOODS and EUROFOODS, I would ask them to review their objectives and take heed of the costs of achieving them. The new UK data bank as yet has taken no account of these organisations since nothing has so far emerged from them that requires firm action. Any changes arising from international bodies will inevitably lead us to considerations of cost/benefit, international obligations and agreements and lead to political discussions. The achievement of a truly international nutrient databank seems a long way ahead; commonality in the overall presentation of data is certainly a lesser task than merger but even so does not appear to be required by many. In the meantime we can all benefit by exchanges between databank users and producers at all levels.

My thanks go to nutritionists Drs David Buss and Hazel Tyler for their patience and help in the production of this paper, to Hamish Kidd and Dr Ashe Kabi of the Royal Society of Chemistry, London, to Dr Piet Van Straten of Unilever, Netherlands and Dr Anders Moller of The National Food Institute (Denmark) for many hours of fruitful and helpful discussions on nutrient databanks.