INFOODS Overview 1988

John C. Klensin

INFOODS Secretariat
Room 20A-226, MIT, Cambridge, MA 02139

INFOODS is now concluding its fourth year, the final year under its initial funding. This paper will concentrate primarily on some recent progress; however, we will first review the overall goals and organization of INFOODS. More detail on project goals and organization can be found in the references listed at the end of the paper.

INFOODS was organized in the winter of 1983, as a project of the United Nations University. Primary initial funding was obtained from the National Cancer Institute, supplemented by funds from the US Food and Drug Administration and the National Heart, Lung, and Blood Institute. Additional funds for specific projects have been obtained from the US Department of Agriculture and several foundations and multilateral agencies.

Primary Objectives

The primary goal of INFOODS has been to improve the quality, use, and interchange of data about the nutrient composition of human food. Since nutrient composition work goes on all over the world, in industry as well as in governments, and there is less agreement than one might expect about how it should be done, this has required two different types of effort. One has included the organizations and people who develop and use the data, and improving communications among them. This effort has included making some important information and practices more generally known and available, since one of the difficulties of the field is that procedures developed or discovered in one country or center may not become known to others, much less comparatively evaluated. The other effort has focused on some new scientific work, not primarily in the chemistry of foods, but in the mechanisms needed to support convenient and accurate exchange of nutrient composition databases among countries and centers. Such interchange requires not only the - fairly simple procedures for transferring the data, but the more complex procedures for encoding data and for describing foods and identifying nutrients and auxiliary data. Since different centers have different conventions, precise identification is a prerequisite to accurate interpretation and use of any data that is received from a "foreign" source.

If we were to summarize the major scientific insight of INFOODS during the last four years in a single sentence, we would say that these problems of description--of determining what data can be compared to what other data and how--are far more complex than we, or perhaps anyone else, had realized. The general problems of naming and describing foods are, as is generally understood, the most difficult and least tractable, but appear not to be the most important relative to the majority of actual scientific use of the data.

Project Organization

To accomplish the rather ambitious objective of promoting improved availability and utilization of food composition data, INFOODS divided its work into four areas and created a separate working and steering committee for each area. Those committees are on Data Quality, chaired by Dr. David Southgate of the United Kingdom; Terminology and Nomenclature, chaired by Professor Stewart Truswell of Australia; Users and Needs, chaired by Professor Carol Windham of Utah State University; and Information Systems, which I chair. Each of these committees was expected to organize itself and meet as necessary, then produce one or more reports and, where appropriate, suggestions for further action.

all countries and cultures can use the same organizations for tables and, in particular, the same food groups and the same terminology. As we look at the way foods are thought of in different cultures, it appears that different cultures may require different organizations. Different, culture-specific, grouping, and data base organization styles seem to be arising spontaneously in various parts of the world. At a minimum, the assumption that it is appropriate for all tables to be intellectually organized according to the same principles is a hypothesis that requires careful examination and testing.

Since these commonly-held beliefs are misconceptions, if one is going to build a database that reflects data from several countries and regions of the world, one must assume large amounts of data, originating from diverse sources, and intended for diverse uses.

Food Identification and Naming

In considering foods and handling data about them, an important consideration is just how to assign names or codes to the foods of interest. Papers at previous nutrient databank conferences have proposed several approaches for both naming and coding methods, or ways of avoiding them. The naming problem is extremely difficult. Of the issues I discuss here, it is most easily confused with the *entire* problem. Time does not permit an in-depth discussion, so I will just quickly review some of the major issues.

In addition to the assumption that a solution to the naming problem will solve all food composition database problems, there has also been a common assumption that a good naming or coding system for one database or purpose will be appropriate for all other databases or purposes; that assumption is rarely true. Often, the better a system, the more dependent it will be on either a small class of uses, a small class of databases, or both. For example, systems should be, and are, designed differently depending on whether they are primarily for use in nutritional counseling or epidemiological studies. Naming systems that must carry the burden of decisions about whether two foods, originating in different tables from different sampling procedures, are "the same" typically require some characteristics that a system for coding foods actually eaten by individuals (for use in nutritional analysis) do not. And, for data originating in other countries, differences in approaches to, and terminology and assumptions about, foods either influence the characteristics of an optimal naming system or require extremely non-specific systems that are not sensitive to national or cultural differences.

Data Formats

Another important belief was that the key problem was table, or data file, organization. The strongest advocates of this belief argued that, if only one could standardize table and file formats--put the same nutrients in each table, and put them in the same order and, perhaps, fields of the same width--all other problems of interchange of data between databases would become insignificant. Even if a standardized format would accomplish that, it is not feasible because different studies need different mixes of nutrients and few, if any, data producers are willing to significantly change their national or internal data organizations and formats in order to accommodate a vague international goal. The diversity of nutrients, and the growing number of nutrients of interest, also makes a standard table format unfeasible: there are just too many to practically organize this way.

At least as important, there seems to be a growing trend toward building multiple, special-purpose data systems, even when they are based on essentially the same data. The continual expansion of this conference's list of systems that utilize USDA-derived data is an example of this trend. In addition to the obvious commercial interests, many of these

systems reflect different perceptions of potential users or uses for the data. Each of those different perceptions may be valid, arguing that different systems, and different types of systems, are appropriate, indeed optimal, for different purposes.

At the same time, there must be some commonality of formats or understanding lest the computer programs of one center be unable to understand data originating in another. One can avoid any commonality only at the cost of having each center that wishes to import data prepare conversion programs tailored to each possible incoming format, and an understanding of that format. People at these conferences have claimed that they have enough problems with the USDA format, which is quite well documented, well understood, and, as these things go, very easy to handle.

The INFOODS strategy for data formats relies on a special interchange format. That format is not designed for day-to-day production use in any particular center, but primarily for the transfer of data between centers in a way that maintains full information (or as much as is available) about the foods and nutrients being reported upon. Such a format must be, as implied above, adaptable to a wide variety of nutrients and other food components and to a variety of description and classification approaches. It should not, in the process, become unwieldy. It must not require that space be left for nutrients unreported, or nutrients that are not yet of interest. It cannot require that changes be made in local or national table formats, or in local databases or procedures, since such changes cannot realistically be expected and are probably not desirable.

The interchange model itself was discussed at the Nutrient Databank Conference in 1986. It depends on the specific "tagging" of all values, and on precise definitions of the tags themselves, using a model based on an extension of the International Standard Generalized Markup Language (see the work of Coombs and his colleagues at Brown for a related discussion). An example, for vitamin C and sodium data about a simple food, would appear as:

```
<comp >
     <vitc> 30 </vitc>
     <na> 0.12 <median> </na>
</comp >
```

This is not as simple and obvious as one might like, nor is it as compact. At the same time, it appears to accommodate the available information reasonably well, and, while not illustrated here, to expand to accommodate as much "metadata"--e.g., detailed sample description of the food analyzed, statistical description of the data values, or even laboratory notes that reflect on the data--as is available. To save space, this example also does not illustrate the data records that contain the food names, codings, and description. Those records are designed to accommodate any naming, coding, or description information that might be available, including national names in national languages, simple food groupings, and detailed classification methods such as the FDA's Factored Food Vocabulary.

A World Database

At least until the various political, administrative, and technical issues are understood, the obvious solution to the incompatibility of databases in different countries and world regions is to put all of the data together into a single database and give everyone access to it. While the approach has a certain obvious appeal and elegance, it is not practical for reasons of several different types.

INFOODS OVERVIEW 1988

The Political

To an increasing degree, countries are becoming proprietary about their data. The data are considered as having value, or as representing the outcome of national investments, and giving them away or letting someone else do so is becoming a complicated problem. At least as important as the cost issue, many data producers wish to maintain at least a modicum of control over the use made of their data and the character of inferences that are drawn from them. Neither of these requirements is well served by an integrated international database.

In addition, if there is to be a single integrated international database, the question of where it should be located, how it should be maintained and updated, where its support comes from, and how decisions are made about it, its use, and how it should be accessed, all become questions that can take on considerable political significance. Such questions often take a long time to resolve; if we must wait for their resolution before beginning more efficient use of each other's data, the wait may be quite long.

The Administrative

If a world database were created, and appropriate arrangements for access worked out, many problems remain. There will be important continuing questions about, for example, how to keep track of changes in various national tables and when those changes should be integrated into the international database. Unless the data from a particular country will be requested very frequently, it may be easier to obtain a new copy of those data when they are asked for than to determine if the version in the international database is up to date (indeed, in many cases, the only practical way to determine if data are current may require obtaining a new version of those data).

The Technical

An international database that must retain not only the data values but a great deal of "metadata"--information about food and sample description, about statistical properties of samples, about specific laboratory conditions that might affect results or representativeness, and so forth--is a very large database, with all of the attendant problems. It would require more computer facilities and support than any (national or regional) subset of that potential database. And, because it is large, it is going to be less efficient for almost any problem-specific use than a database devoted to, and optimized for, that problem. A unified international database as an archival repository would be expensive to develop and more expensive to maintain, raising the issues cited above. It is also unnecessary for any purpose we have been able to identify, since distributed approaches are possible and more convenient and economical. For direct use in analysis or calculations, as contrasted to archival use, such a database becomes less appropriate because of its very size and the facilities needed to deal with that size. The early INFOODS notion of someone sitting in a jungle with a laptop terminal and small satellite dish, and pointing the latter toward a satellite link to "the" international database, is neither realistic nor reasonable: capturing the relevant data and carrying them with the laptop machine is less expensive, more reliable, and a great deal faster in terms of use of the data.

As a consequence of these difficulties, it would appear to make more sense to let the producers of data maintain them, rather than trying to create a central, international archive. To do that, we need two things: a good directory of what data are available and

where, and mechanisms to transfer data smoothly, efficiently, and in comprehensible form. It appears that, contrary to the popular assumption, a single, comprehensive, and integrated international database is neither necessary nor even desirable.

The INFOODS Committees and Their Work

The Data Quality committee has focused its work on a new manual of methods of data analysis and handling for food composition work. The technical work for that manual, being written by Dr. Southgate and Dr. Heather Greenfeld of Australia, has been completed. The manual is being edited as this paper is being written, and it should be available later this year.

The Terminology and Nomenclature committee has developed a series of reports on the description of foods--structured description, rather than assigning names--which are available, in preliminary form, from the Secretariat. Summaries of the main results of those reports are awaiting publication. That committee has also worked with the Secretariat to produce a first draft of some generic keywords for food identification: that work is, deliberately, less adequate than any good, problem-specific, food coding or naming scheme, but is likely to be helpful, especially in dealing with foods from little-understood cuisines, when such naming or coding is not available.

The Users and Needs committee was asked to produce a survey of the users and uses of food composition data, and to discuss the requirements of future work. Its work has produced a book on the subject, and the committee has been inactive since that time.

The Information Systems Committee began its work as a facilitator of the use of computer technology, and information systems technology more generally, for the rest of the project, as well as working on its primary role of arranging facilities and for the exchange of food composition data and databases between regions of the world. It started with a strong hypothesis that most of the problems encountered with food composition data needed scientific solutions; that the computer could be of use only when the problems were solved, or, at least, better understood. That hypothesis has proven true, but work on the exchange of data has identified and clarified a series of problems that were not previously generally understood. The Information Systems activity has found itself identifying, and clarifying the character of, certain problems--such as the specific identification of "nutrients", discussed below--in working with the Data Quality and Terminology committees to develop specific proposals and standards.

Nutrient Names and Nutrient Identification

We have also encountered a general assumption that the identification of nutrients is fairly easy. As we pointed out in the paper Dr. Rand gave last year, this is not the case. Different tables use the same "nutrient name" to describe rather different things. In some cases, we know that direct comparison of the values is incorrect and that the values are different in biologically significant ways. For example, a "vitamin A" value that represents retinol only should not be compared to a "vitamin A" value that represents the sum of retinol and a series of carotenes and carotenoids. In other areas, such as several available measures of dietary fibre, the biological significance of the differences in values is not clear (and is still a subject of research).

Other researchers have assumed that, because the analysis procedures are different or the underlying chemicals are different, we should not identify "nutrients" in tables at all, but only chemically clearly-distinguished food components, distinguishing further among the methods used to obtain them. While this approach has some appeal, it is not realistic for

dealing with foods, nutrients, and their impact on human health and well-being.

Since it appears that neither of these assumptions is completely accurate, we must use a mechanism for identifying nutrients very precisely, distinguishing among methods only when the methods would make a difference in the expected values that end up in tables or databases. In other words, if two analyses or calculations include different chemical components, and thereby produce different numbers, we should consider them as associated with different "nutrients". An adequate system for nutrient identification did not exist before INFOODS began its work, so we have developed one, with the assistance of many reviews and comments from around the world. Either the new INFOODS system, or some other one, should, ideally, be internationally standardized. In addition to the value of a nutrient identification system in data interchange, such a system should provide a shorthand, and very precise, way of identifying what types of analyses and processes have been chosen for a particular table in the introductions to the tables.

INFOODS has adopted exactly that approach. A list has been developed that reflects every nutrient that appears in a major food composition table or database in the world (and many minor ones). The list distinguishes among methods of determining a particular "nutrient" when those methods might reasonably be expected to produce different results for the same foods. Abbreviated "tagnames" have been associated with each of the nutrients and nutrient determination methods so that they can be easily recorded and used in interchange. The list has been extensively circulated and many comments have been received, leading to some modifications. The current version is available on request.

It is important to note one difference between this list and the more general practice of recommending or specifying particular methods within a country or for particular uses. The purpose of the INFOODS list is to identify what has been done, not to make judgments about what should have been done. It includes methods that have been discredited or superseded but from which data may still be available. By including those methods, it becomes possible for the receiver of data to decide whether or not those data are worth using (they may be good enough for some purposes, or nothing else may be available), rather than having that judgement made independently of an understanding of the problem and the degree of accuracy or precision that is required.

The Directory Problem

If, as is suggested above, we are to create and maintain an international directory of food composition data, we must first be very clear about exactly what type of directory we wish to build and what is to be included in it. For example, a directory of food compositiontables, such as the current INFOODS directory and the older FAO directories, is different - certainly in content and probably in optimal organization--from a directory of food composition databases. Neither is the same as a directory of where food composition data might be located, since that would, ideally, include a collection of references to journal articles and monographs that contained data values. While those types of directories are at least related, a directory of food composition data systems, focusing on the data as they appear through accessing and computational software, is probably a different type of structure entirely.

Directories of Tables

A directory of food composition tables, such as the FAO directories or the INFOODS one, is the easiest of these types of directories to organize. Such a directory does not require very

much information to be useful. Minimally, it appears to require the following information about each table:

* The region or country for which the data apply.
* The author or source of the table or data.
* The publication information for the table, including its title, publisher, and how it may be obtained.
* The publication date and, if different, the date(s) reflected by the data (e.g., for foods that change over time, knowing the times at which sampling occurred is vital to being able to interpret the data).
* The size or length of the table, preferably in terms of the number of foods and nutrients represented, but in pages as an alternative.

Directories of Databases

An index of databases of course requires a significant amount of additional information. That information would probably include:

* Clear distinctions between databases and data systems (with associated accessing or computational software).
* The foods or food groups included in the database.
* The nutrients included, and how precisely they are measured and to what standard.
* The systems for classifying or naming foods which are represented in the database.
* The types and extent of data description that are present and how they are organized.
* The type of database organization and its implications for the user. In particular, what software, or types of software, are required to use or access the data? And is the INFOODS Interchange Format, or some other convenient format for data interchange, supported? Or will special software have to be built to convert from one format to another?

The critical questions, of course, are "does this database contain what I am looking for", "is it appropriate for my needs", and "how can it be made available?". The directory must permit, at least, most of the answers to these questions before the user incurs the time and expense needed to actually retrieve and examine the data themselves. As we gain experience, the list above is likely to change somewhat, or to become more precise within these categories.

Data Quality - A multidimensional issue

To be of greatest use, a directory should not be just a listing (of whatever it lists) but should provide some information about the relative quality and usefulness of the data in the table or database referenced. There has been a good deal of interest in data quality measures during the last several years, concentrating on a single data quality scale ranging for example, from "high" to "low". INFOODS research has led us to conclude that a single quality measure is inappropriate and unattainable: the quality issue is multidimensional, with different combinations of dimensions being relatively more or less important for different uses or users. Our work in this area is barely started: we can claim to be near to understanding the problems, but not to being ready to make a proposal about a solution. Some of the important dimensions are:

INFOODS OVERVIEW 1988

The "Chemical" Dimension

This dimension is one of those usually discussed as being "data quality". In one common terminology, it involves a scale as to whether the data are "analyzed", "imputed", "copied", "copied and adjusted", or "guessed at", but there are many variations on this theme. For "analyzed" values, the quality, reliability, and reproducibility of the method used is also important here, or may be part of another dimension.

The "Sample" Dimension

There may not be a single sample dimension but the issues that must be addressed include:

* How the foods to be analyzed are selected in the field
* What procedures and methods are used to handle the foods on the way to the laboratory
* How the samples are drawn and handled in the laboratory
* How representative the results are relative to foods encountered by the consumer or eater

There appear to be a number of other dimensions, but neither time nor our understanding permits a complete listing and explanation. The following should at least be examined as contributors to "data quality" and "data usability":

* Reporting quality
* Identity of foods

INFOODS Status today - The Conclusion of the First Four Years

To review where we are today, INFOODS has accomplishments or significant progress in three major groups of areas.

Organizational and Scientific Communications

NEWSLETTER

We have been publishing a newsletter on a roughly quarterly schedule, which many of you have received. The newsletter identifies topics and meetings of interest (of which we are aware), newly-published or available food composition tables or databases, and INFOODS-related information. It has also been used as a forum in which new developments or ideas in the management of food composition data can be presented. Future publication of the newsletter is contingent on our finding financial support for it.

JOURNAL

The newsletter is quite informal. It supplements a new journal, the Journal of Food Composition and Analysis, sponsored by the UNU through INFOODS. Several issues of that journal have now been published by Academic Press.

USERS AND NEEDS

As mentioned above, the Users and Needs committee has produced a book, Food Composition Data: A User's Perspective. That book is available from the United Nations University office in Tokyo.

J.C. KLENSIN

DIRECTORY OF FOOD COMPOSITION TABLES

A preliminary version of the second edition of this directory is available for restricted distribution. The complete version of that edition is being formatted and printed and should be ready for distribution in August of this year.

GUIDELINES FOR ANALYSIS AND DATA

This book is in the final editing stages.

GUIDELINES FOR TABLE CONSTRUCTION

This book is awaiting last internal review and edit before being circulated to meeting participants and other reviewers.

DATA COMMUNICATIONS AND INFORMATION SYSTEMS

A detailed description of the interchange model itself, consolidating a number of working papers and notes, is being prepared and drafts of some chapters have been circulated. A paper on methods for describing foods, as discussed above, has been prepared and is awaiting publication. It is supplemented by lists of descriptive terms and a list of questions that might be suggestive as part of the process of developing or reviewing food descriptions. Also as discussed above, the "tagnames" for nutrient identification have been reviewed in three separate versions. A fourth version, incorporating the comments on the previous versions, new sets of food table references, and an appendix that provides hints for distinguishing among different variations on the same nutrient, is in final preparation and should be available before September.

A document has been prepared that describes the requirements for, and operation of, regional food data centers; that document was reviewed at the INFOODS policy. committee meeting in Budapest during late 1986 and is available from the Secretariat. Finally, preliminary working drafts have been prepared on a general-purpose food classification system; little additional work in that area is anticipated at present due to the press of other work and interests.

In many cases, these work items are leading to recommendations about how food composition database developers should identify data and organize databases to facilitate international interchange and comparison of data values among tables. We are working with international organizations, especially FAO, to create formal standards from these recommendations when that is appropriate.

The first four years of INFOODS have been characterized, we believe, by a great many accomplishments that are especially significant when considered against the background of an ill-defined field that can be described as dominated by a great many misconceptions. Those misconceptions stem from a series of problems not unlike those of the proverbial blind men trying to describe an elephant: different groups, working with different aspects of food composition data, have tended to see their problems, views, or proposed solutions as the entire problem or the whole solution. We have discovered that this is not the case, and have begun the work of describing the entire elephant and, to some degree, to minister to

its ills. We hope that we can come back a year from now and report even more, and more specific, accomplishments.

DOCUMENTS AND MATERIALS REFERENCED

Coombs, James H., Allen H. Renear, and Steven J. DeRose, "Markup Systems and the Future of Scholarly Text Processing", Communications of the ACM, 30, 11 (November 1987), pp. 933-947.

Greenfeld, H. and D. A. T. Southgate, Guidelines for the Production, Management, and Use of Food Composition Data: An INFOODS Manual. To be published, 1988.

Klensin, John C., Diane Feskanich, Victor, Lin A. Stewart Truswell, and David A. T. Southgate, "Identification of Food Components for INFOODS Data Interchange", INFOODS Working Paper INFOODS/IS N40, August 1988.

Rand, William M., Carol T. Windham, Bonita W. Wise, and Vernon R. Young, eds., Food Composition Data.: A User's Perspective, (Food and Nutrition Bulletin Supplement 12), Tokyo: The United Nations University, 1987.

Rand, William M., Jean A. T. Pennington, SuzaIme P. Murphy, and John C. Klensin, Compiling Data for Food Composition Databases. To be published, 1988.

Truswell, A. S., D. Bateson K. Madafiglio, J. A. T. Pennington, W. M. Rand, and J. C. Klensin, "Facets for the Description of Foods: INFOODS Guide to the Aspects of Foods Which Should Be Considered when Describing Them for a Food Composition Database". Pre-publication draft available from INFOODS Secretariat.