John C. Klensin

INFOODS Food Composition Data Interchange Handbook

INFOODS
Food Composition Data Interchange Handbook

JOHN C. KLENSIN

United Nations University Press

The International Food Data Systems Project (INFOODS) is a comprehensive effort, begun within the United Nations University's Food and Nutrition Programme, to improve data on the nutrient composition of foods from all parts of the world, with the goal of ensuring that eventually adequate and reliable data can be obtained and interpreted properly worldwide. At present in many cases such data do not exist or are incomplete, incompatible, and inaccessible.

This volume is the fourth in a series that provides information and guidelines about requirements for food composition data, the identification of nutrient and non-nutrient components of foods, the computer representation and accurate interchange of food composition data, and on the organization, compilation, and content of food composition tables and data bases. It presents the structure and rules for moving data files between countries and regional organizations in a way that preserves all of the information available. The approach also alerts the developers of data bases about potential areas in which ambiguities are likely and special care should be taken and identifies some mechanisms for improvement of overall nutrient data base quality.

The United Nations University
Toho Seimei Building, 15-1 Shibuya 2-chome, Shibuya-ku, Tokyo 150, Japan
Tel: +81 3 3499-2811 Fax: +81 3 3499-2828
Telex: 25442 UNATUNIV J Cable: UNATUNIV TOKYO

Printed in Hong Kong

WHTR-16/UNUP-774
ISBN 92-808-0774-9
United Nations Sales No. E.91.III.A.6
03000 P

John C. Klensin is Principal Research Scientist in the Department of Architecture, Project Coordinator for INFOODS, and Director of the INFOODS Secretariat at the Massachusetts Institute of Technology, Cambridge, Massachusetts.

Many people made significant contributions to the development of the INFOODS Interchange System. In particular, Dr. David Peterson and Ms. Roselyn M. Romberg contributed many ideas and some text to earlier versions of this document. Comments from Mr. Craig Franklin and Dr. Zita Wenzel, as well as Dr. Peterson, helped to determine the special forms of using the SGML Standard. Professor Vernon R. Young, Dr. Lenore Arab-Kohlmeier, Ms. Diane Feskanich, and several others provided feedback at critical times about the relationship of the evolving system to possible practice with nutritional data and their use. Anders Møller, Lena Bergström, Brucy Gray, Pam Verdier, and the New Zealand Division of Scientific and Industrial Research provided data against which the conversion models could be effectively tested, some of which is incorporated, with permission, in the examples. The material on the description of data files being interchanged and on the description and classification of foods is a formalization of material, some of it still unpublished, developed by the INFOODS Committee on Terminology and Nomenclature, headed by Professor Stewart Truswell. The data description section derives from several discussions and position papers about the basic character of ideal descriptive statistics for small samples and unknown distributions with Dr. Ree Dawson and Professor William M Rand, both of whom also made frequent and helpful comments about other parts of the manuscript and the working documents that were the foundation for it. Finally, the system was presented in technical detail at a special Oceaniafoods technical workshop on food composition data base organization and interchange. The participants in that workshop provided invaluable feedback on both technical issues and on the clarity of some of the concepts.

Any work of this type is ultimately a synthesis of many ideas and concepts. Much of the credit should go to those who contributed; the blame for the interpretations rests, as always, with the author.

Contents

PART I
INTRODUCTION AND OVERVIEW

1
Introduction to the Interchange System

AN INTERCHANGE SYSTEM FOR FOOD COMPOSITION DATA

A major goal of INFOODS has been the development of easy and accurate interchange of food composition data among countries and regional organizations. Such data exchange will obviate the perceived need for a single international data centre which holds all of the world's data, replacing it with distributed arrangements in which most data are held by their compilers, or by regional data centres operated by organizations of which data compilers or owners are members, until the data are actually needed.

It is not sufficient merely to move data back and forth. Food composition data are complex and often are, or should be, accompanied by extensive description of the foods being reported upon and the methods of analysis used. It has become clear in the last few years that the introductory material in a printed table may be nearly as important as the data values (see, for example, Arab et al. [1]). The need for such description and explanation arises through the necessity of comparing data from widely differing cultures. Not all food composition tables and data bases have the same level of description, however, and the informal text of an introduction is not the best way to communicate the information that is available, especially if it is to be processed automatically (e.g., by a computer), rather than simply read by trained scientists.

Other distinctions have been noted about various types of tables and data bases Some data bases are oriented toward end users, others for national reference purposes, and still others are the fundamental collections of laboratory-level data before aggregation [24]. An effective interchange mechanism must be able to handle any of these types of data, without obscuring the differences in the types of information contained in each

As one examines international data interchange, it becomes clear that the primary criterion for designing and evaluating a data interchange system is that it preserve whatever information actually is available, without forcing the data supplier to provide any more information than is known or imposing any more burden than is absolutely necessary. It would not be reasonable to try to require data suppliers to supply information which they do not know, or do not normally keep for their own purposes. Similarly, while in an ideal situation everyone might do things in the same way, the interchange system must be able to accommodate methods of reporting and data organization that some scientists might consider inappropriate. The inclusion of a way of expressing a particular concept in this document is therefore not necessarily a recommendation of that concept. Indeed, in a few cases, the text recommends against styles of data presentation and identification for which provisions are nonetheless made Because identical and accurate sampling and analytic procedures, food selection, data description, and reporting are unlikely to ever occur in all tables, successful and meaningful exchange of food composition data has necessitated developing new conventions and technologies to organize and identify the many and varied components of these data.

Accurate comparison of data values requires very precise identification of how the values were derived and what they mean. When existing food composition tables and data bases are considered without their sometimes detailed introductions and appendices, there are often major ambiguities concerning the exact identification of foods, nutrients, units, and analytic and sampling methods More careful comparison of food composition tables shows that different tables provide information about different nutrients, different types of foods, and different amounts and types of supporting information about samples, quality, recipes, and so on.

While any approach must accommodate the data that exist, the nutrient composition field continues to evolve New food coding systems are introduced frequently, and changing hypotheses about the relationship between foods and health result in the introduction of nutrients that were not previously considered interesting into tables and data bases. If an interchange arrangement is to be useful for more than a few years, it must be "extensible", i.e., it must provide for new terminology, technology, and areas of interest to be defined and added to the system without compromising existing files and programs.

The differences in values and the ambiguities of data and food identification inherent in existing food composition data require that any interchange model operate on the assumption that actual tables and data bases cannot be expected to conform to a single standard or format. The interchange strategy must be descriptive of what decisions have been made about foods, food classification, nutrients, chemistry, or description and how those decisions have been carried out. At the same time, as suggested above, it cannot be dominated by norms about the "right" way to do things: even questionable data, poorly organized, may be more useful than no data at all, especially if the nature of the problems can be carefully identified and understood

Partially as a result of the fact that particular data may be acceptable for some purposes and not for others, another goal of the interchange system is to permit tracing the flow of values, through copying (borrowing) or calculation, from one table to another and, more important, to be able to trace and assign responsibility for those values. *All of the requirements for information that must be supplied with interchange files are the result of this tracking requirement.*

To permit data interchange without loss of quality, and to encourage improvements in quality, data description, and data definition, INFOODS has designed a system of regional data centres and has developed an "interchange system" by which whatever data exist and are of interest can be transferred among regional centres with precise identification of values and without any loss of information. The interchange system is both a model of how data can be transported between regional centres and a data interchange format definition. As the latter, it is derived from principles of "generic markup" which are becoming increasingly important in the processing and exchange of textual documents. The standard for generic markup is specified in widely adopted international standards based on an International Organization for Standardization document, ISO 8879 [53]. Using generic markup has several special attractions, including its growing availability, the ability for people to directly inspect the format and content of the files, and the lack of dependence on any particular medium or data-transport arrangement. The other alternatives which are possible in principle were systematically eliminated as infeasible or too restrictive [55].

The interchange system will be used internationally, to facilitate exchanging data among countries and regions of the world. As with other INFOODS work, the interchange system uses existing international standards whenever possible, even when the invention of a nearly equivalent set of conventions specific to food composition data might result in short-term convenience or compactness. For example, provision is made for expressing food names in national languages and character sets where necessary, but only when consistent international standards for those character sets have been established.

THE REGIONAL DATA MODEL

While the details are not discussed in this manual, operating regional data centres, affiliated with INFOODS, are assumed as part of the interchange system. Those data centres act as a focus for food composition data base activity in their regions of the world and as the host for data interchange activity When data are needed, for example, in most circumstances the user requiring the data would contact his or her own regional data centre, which would make arrangements to obtain them from a distant regional data centre, which might, in turn, obtain them from an organization within its region. The interchange mechanisms described in this manual are required only for use between regions. While they may be suitable for use between a regional data centre and data providers or users within its region, and may also be suitable for the ongoing storage of some reference or archival data bases, regions are free to work out their own arrangements for intra-regional communications and data interchange. A region that has specified its own data interchange formats and arrangements will presumably provide the capability to convert between the formats and conventions specified in this manual and its own formats at its regional data centre.

A regional data centre will typically be operated as part of an INFOODS regional liaison group, but this is not a requirement; either could exist independently of the other, and the term "regional data centre" is used instead of "regional centre" to stress this distinction. In principle, the regional data centre for a particular region need not even be located in that region, although it would usually be desirable for it to be.

In addition to acting as a focus for data interchange activities for its region, a regional data centre is expected to act as a registrar of international food record identifiers for the associated region (see page 66), maintain current lists of interchange system tags and other identifiers, and keep records of tables and data bases originating in the region. It may also maintain some data locally, either from within the region (for easy export or as part of regional support functions) or from outside the region but frequently needed within it. In either of these cases, the regional data centre is expected to make special provision to ensure that its copies of data sets are kept up-to-date or that they are discarded when they are no longer current.

THE INTERCHANGE SYSTEM AS A CONCEPTUAL DATA BASE MODEL

While the principal design goal for the interchange system is information-preserving exchange of data among regional centres, its provisions for precise identification of nutrients and other food components, detailed recording of varying amounts of data about each nutrient and descriptions of those values, and ability to accommodate multiple coding, classification, and description systems may make it appropriate for national or regional use for archival and perhaps reference data bases. INFOODS has not made a specific recommendation that it should be used this way, but if the character of the data and description associated with a data base creates difficulties in using conventional data base systems with statistical or scientific data [4, 18, 27] the architecture of the interchange system, and software developed to handle it, might be considered as an alternative.

THE CONCEPT OF AUTHORITY

Food composition data, like most other scientific data, are rarely "true" or "false" in any absolute sense. Instead, the data values, the choice of foods, the decisions about whether two samples represent the same food, or a set of samples adequately represent some particular food, all represent scientific choices, not completely deterministic outcomes of perfect processes. In particular, it is possible, indeed likely, that different but equally skilled scientists would make different decisions, especially under different circumstances or assumptions about the user population and its needs.

As part of the important goal of preserving to the greatest extent possible all of the information about data being stored or transferred, an interchange system must move beyond traditional styles of exchanging only individual values in two important ways:

It must encourage asking, answering, and documenting questions about what person or organization made a particular decision and will take responsibility for it. For example, in Chapter 4, a restriction is imposed that a single interchange file must have only one "source". This does not imply that all the data must come originally out of the same laboratory, or even the same country. Instead, it recognizes that the activity of putting together a data base involves editorial and scientific judgment, rather than mechanical concatenation of values. This is especially true when the data are derived from multiple sources. If nothing else, someone must conclude that combining the various values and considering the combination "one data base" makes sense. As soon as that decision is made, we have a new data base, containing new information—the decision itself—not just a combination of other data bases. And that implies a new, separate (and single) source.

Similar issues apply at the level of "individual foods". As discussed in Chapter 5, each collection of data associated with "a food" is associated with a food record identifier. A data base may contain multiple records for a given food, with different sets of values. If it does, each of these records will have a different food record identifier. The decision about whether a single food should have one or several food records is made by the table compiler. The interchange system imposes only two rules: (i) If previously published and identified data for an entire food (i.e., a single food record) are copied together, the food record identifier must be the same as the corresponding one in the original table or data base. That is, the authority and responsibility for the integrity of the data rests primarily with the compiler of the original table or data base (but not the decision to include the data in the particular new data base). (ii) By contrast, if a food record is assembled from multiple sources—e.g., proximates and vitamins from one country and minerals from another—several key scientific decisions go into the compilation and combination process, and a new food record identifier is assigned to the newly created food record.

Biological variability, variations in recipes, and many other factors contribute to there rarely, if ever, being a simple, firm value for the amount of any component in any food. Instead, the values typically represent some estimate of a particular parameter or other property of a statistical distribution. The interchange system provides extensive mechanisms for describing the distributions, and knowledge of and beliefs about them, in addition to simple values, or values and standard deviations or errors of estimate heretofore prevalent. These facilities are discussed in Chapter 6. While significant use of these facilities is not anticipated during the first years of data interchange, it is intended that they should provide a model for structuring more detailed information. That information should gradually become available as sophistication increases about data management and reporting within the food composition data user and supplier communities.

THE ROLE OF THIS MANUAL

This manual defines the organizing principles and formats of the interchange system—the model by which data about food composition can be transferred from one facility (typically a regional centre) to another while structuring and preserving whatever information may be available. It also specifies the ways in which the interchange system and its elements can be extended to account for changes in scientific conventions or knowledge without requiring data bases to be changed or programs to be rewritten if the changes are not important relative to the content or users of those particular data bases or programs.

The interchange system, of which an overview appears in the next chapter, depends on these principles and on conventions about the syntax in which textual and numerical values are written. As with conventional textual use of generic markup, the essential syntax uses a collection of carefully-defined "elements" which, in turn, are identified by a collection of specifically-defined "generic identifiers".* Generic identifiers are predefined word-like strings of characters used to distinguish one element type from another.

More precise definitions of these terms, and examples of how they apply, appear in the chapters that follow. Later chapters specify those elements which are part of the interchange structure itself; the structure of elements used to describe the origins of, and responsibility for, an interchange file; foods and the properties of data. While the structure of elements that contain data values about the quantities of individual components present in foods is specified here, the generic identifiers for the food components themselves are specified elsewhere, primarily in the food component

* INFOODS materials that discuss aspects of data interchange for other audiences often use the term "tagname" to refer to the significant part of a generic identifier. That term is avoided in this manual to preclude the confusion that arises when the "name" of an element and its written representation differ.

identification listing [17]. The information in that book may be needed for in-depth understanding of some of the examples that appear here. With the exception of a few areas for which specific generic identifiers have not been assigned at the time of publication, every element that appears in this manual is described either in the reference sections or in the food component identification listing.

The general model of the interchange system is applicable to a great deal of food-related data which are not yet defined for use with it. Decisions to limit the extent of what to define have been conditioned by finite resources, the focus of the initial INFOODS mandate, and lack of clarity either of the needs or of the appropriate solutions. When additional elements of these are needed, working papers that begin to explore their development will be commissioned. These as yet unneeded areas and definitions include the use of national character sets for other than names of food, listing of recipes for mixed dishes, listing of food economics values (e.g , food balance data or food prices), and listing of food components that are not normally considered nutrients (e g., food additives and contaminants)

PURPOSE AND AUDIENCE

This manual provides sufficient information about the interchange system to permit programs to be correctly written that will produce and interpret interchange files. Readers who are only concerned about a general introduction to the interchange system should concentrate on Part I, reading quickly through the balance with the confidence that most of the details are not important to them. Nonetheless, this is a technical document, and some terminology is used in very precise ways. The glossary contains all such terms, and should be consulted when there is doubt about whether a word is being used casually or with some special meaning.

Finally, this manual does not discuss the particular methods of transporting an interchange file from one location to another. The interchange system is designed to be insensitive to the choice of media (e.g., magnetic tapes or floppy diskettes) or transport mechanisms (e.g., computer networks or the post), depending only on a specially-delimited "interchange file". Since an interchange file consists only of text, it can be transported by any medium—including file transfer or electronic mail in computer networks; magnetic or optical recording on tapes, disks, paper, or diskettes; or even such older media as punched cards or paper tape—so long as the medium is able to transport eight-bit characters accurately. If elements that can contain "national characters" are removed from the file before it is sent or ignored when it is received, transmission with media that can process only seven-bit characters, or even low-quality computer printouts and telefax transmission and subsequent optical scanning are feasible. The only requirement is that the interchange file must be clearly separable from other information, a requirement that the file definition itself enforces. Sender and receiver should, of course, reach agreement about the media and mechanisms to be used before data are actually transmitted. Conventions about media and mechanisms for interchange among INFOODS regional data centres will be developed depending on the facilities available at those centres.

2
Technical Overview

BASIC TERMS AND DEFINITIONS

The structure of an interchange file is described in terms of *elements*, or precisely identified blocks of data.* The element is the basic "building block" of an interchange file, and serves to identify and contain the actual data being exchanged. Elements provide a structure for the data which is logically ordered for machines and relatively easy to follow for human beings. A typical element might be:

<NA> 5 </NA>

Elements are identified by *tags,* which identify and surround *contents*. In the example above, <NA> and </NA> are the tags which surround the content "5". Some elements use only a single tag, and are delimited by the next tag in sequence, whatever it might be. For example:

<date> 1983.11.04

Here, the content is the string "1983.11.04" in ISO standard date format [41], meaning "4 November 1983", the actual data content of the element. Contents may be *data values* (i.e., numerals or unrestricted strings of text), *keywords* (i.e., special values from a restricted list), other elements, or a combination of values, keywords, and elements. Elements that occur within other elements are said to be *subsidiary* or *nested,* and the term *immediate* is used to denote direct nesting, without intermediate elements, when the distinction is important. The following example, a brief but typical food component or

* The terminology and structure defined here are derived from, and consistent with SGML (ISO 8879) [53], discussed in more detail below.

<comp> element, contains a combination of data values, keywords, and nested elements and illustrates these concepts:

<comp>
<VITC> 30 </VITC> <NA> 0.12 <unit/> MMOL </unit/> </NA>
</comp>

In this example, the food component element consists of two tags, <comp> and </comp>, called the *start-tag* and *end-tag* respectively, and a content of two nested elements. The first subsidiary element is the vitamin C element, whose tags are <VITC> and </VITC> and whose content is the actual data value "30 milligrams per 100 grams edible portion of food" (the units are specified as the default in the definition of the tag associated with the identified food component [17]). The second subsidiary element is the sodium element, whose tags are <NA> and </NA> and whose content consists of a value and a subsidiary element which specifies the unit of measure. The unit element's tags are <unit/> and </unit/> and its content is the keyword "MMOL", which stands for "millimoles". The <VITC> and <NA> elements are immediately subsidiary to <comp>. <Unit/> is immediately subsidiary to <NA>, subsidiary (but not immediately subsidiary) to <comp>, and not subsidiary to <VITC> at all. When it is clear from context which is meant, as in the case above, the start-tag is referred to as if it were the element. For example, in the previous sentence it would be more precise to say "The <unit/> element is immediately subsidiary to the <NA> element...".

Spaces before and after elements and line breaks are ignored in the interchange system. Hence the example above could be written all on one line, or with the sodium and vitamin C elements on separate lines, and so forth.

STRUCTURE OF AN INTERCHANGE FILE

In order to permit processors for interchange files to interpret them accurately and efficiently, interchange files must adhere to certain structural conventions. Consistent structure for all interchange files facilitates ease of use and interpretation of the data, both by people and by machines.

Every interchange file contains a single <infoods 85> element. Other types of information, such as data about the transport medium (e.g., magnetic tape density), electronic mail headers, telex information, mailing addresses, and informal text associated with the transportation of the file may surround but are not part of an interchange file.

The <infoods 85> start-tag is the only tag in the interchange system which requires an "attribute" indicating the version of the interchange system in use, in this case the version dating from 1985. The first tag of an interchange file must appear, therefore, as <infoods 85> and the last one must be </infoods>.

The <infoods 85> element's content is made up of two or more subsidiary elements, appearing in this order:

- a <header> element,
- an optional <dflt> element, and
- one or more <food> elements.

The <header> element identifies the sender and the source of the data. The <dflt> element identifies defaults which apply to the entire data file, such as weights and measures. The <food> element classifies the specific food, identifies any relevant measures, and supplies the relevant nutrient composition data for the food. The structure of an interchange file is therefore:

```
<infoods 85>
 <header>
   source and sender elements
 </header>
 <dflt> default elements </dflt>
 <food>
   <classif>
     <ifri> food record identifier </ifri>
     other classification elements
   </classif>
   <fddflt> per-food default elements </fddflt>
   <comp> food component data elements </comp>
   <drvd-comp> derived food component elements </drvd-comp>
  </food>
 other food elements, starting in <food> and ending in </food>
</infoods>
```

While the <header> element is supplied once and not repeated, and the <dflt> is either omitted or supplied once, the first <food> element would ordinarily be followed by additional <food> elements, since it would be rare to transmit information about only a single food. All interchange files must adhere to this structure as outlined in the example above. (Again, line breaks are ignored in actual interchange; they are used in this book merely to enhance readability.)

OVERVIEW OF THE INTERCHANGE FILE PRIMARY ELEMENTS AND ELEMENT GROUPS

The Header

The <header> element of an interchange file provides information about the sender of the file and the source of the data. This information is critical in identifying the data for interpretation and for archival and tracking purposes The <header> element is composed of two subsidiary elements, the <sender> element and the <source> element, each of which is composed of a number of required elements with several additional elements optional. The list of <header> elements and their definitions is inspired by the work of the INFOODS Committee on Terminology [33].

The Sender Subsidiary Element

The <sender> element of the header is composed of elements that identify the sender of the interchange file. This is the person or organization responsible for preparing the file at hand for transmission, *not* the person or organization responsible for the data values The information in this element must be available to the receiver or user of the file to permit contacting the right person if there are problems with the organization of the data.

Required elements include those for name, organization, address, location or country of sender, postal code, and date of transmission of the file. While some of the information is redundant, the repetitions are important for sorting and classification purposes. Optional elements include those for additional information which is useful but not critical, such as the sender's title, electronic mail address, international telephone numbers (voice and fax), telex number, and cable code.

The Source Subsidiary Element

The <source> element of the header is composed of elements which identify the source of the data—typically a table or data base and compiler—being transmitted in interchange form. *Only one data source is allowed per interchange file.* Possible data sources may include food tables and other publications, nutrient data bases, laboratories, and so on. Optional elements include the address of the analytic lab if the source is a laboratory, the publisher's address for a literature source, or the ISBN number for a book.

The idea of a "source" involves several issues about what foods should be reported, or used, as a single "table" entity. It is most easily understood by analogy to the concept of data for a single food. The realities of chemical analysis and laboratory measurement make it improbable that nutrient values for a single analysis will all be from the same individual food item (e.g., the same apple), nor would we expect values derived from a single apple to have any special merit Instead, one samples, homogenizes, and combines items to construct a laboratory sample [11]. The decision as to which apples are representative of "apple" or even of a particular cultivar and set of growing conditions is a substantive scientific one, and the criteria of "sameness" are neither trivial nor obvious.

While the <sender> element describes the origins of the interchange file, the <source> element describes the origin of the data values themselves. Information provided with <source> might be used to obtain additional scientific information about the data; information provided with <sender> is useful for technical problems with the interchange itself. In addition, <source> is expected to contain the information needed to reference the data in a publication that uses them By contrast, <sender> would provide information for an acknowledgement of someone who had been particularly helpful.

The following is a complete sample <header> element:

```
<header>
 <sender> <date> 1988.06.07
  <fullname> Dr. J. D. Smith <fsnm> Smith
  <orgz> EUROFOODS Regional Centre
    <addr/> Department of Human Nutrition <-> Agricultural University <->
    De Dreijen 12 <->6703 BC Wageningen <-> The Netherlands </addr/>
    <country> NL <postcode> 6703 BC
    <title/> Coordinator of the Laboratory </title/>
    <phone/> +31 83 70 8 25 89 </phone/>
    <telex/> NL 45015 </telex/>
 </sender>
 <source>
   <ref/> Souci, S.W., W. Fachmann, H. Kraut Food Composition
     and Nutrition Tables, 1986/87. Stuttgart: Wissenschaftliche
     Verlagsgesellschaft mbH, 1986.
     <pub/> Wissenschaftliche Verlagsgesellschaft mbH </pub/>
     <isbn> 3-8047-0833-1 </ref/>
   <addr/>
```