13 Jul 2007

Proposal for structure and detail of a EuroFIR standard on food composition data

I: Description of the standard

Prepared by W Becker, I Unwin, J Ireland, A Møller

Contents

1 Introduction

The term “Food composition data” includes all information referring to the description and identification of foods and their content of components. One of the aims of EuroFIR NoE is to develop a standard that will be used as a framework for compiling and disseminating food composition data that are comparable and unambiguous with respect to the identity and description of foods, components and compositional values. Thus the standard should describe procedures for both data management and data interchange.

The present proposal is based on the COST99/Eurofoods recommendations intended for food composition database management and interchange (Schlotke *et al.*, 2000) and, where appropriate, text from that document has been incorporated or adapted in the present document. Experience from other projects, e.g. EPIC/ENDB (Charrondière *et al.*, 2002), BASIS (Gry *et al.*, 2002) and from existing standards (e.g. ISO standard on food safety management systems, ISO 22000:2005) is taken into account. The degrees of detail specified in the EuroFIR food composition recommendations will depend on what is feasible on a European level.

In the Eurofoods recommendations, tables were listed containing the main objects/entities needed to describe food composition data. An entity represents a distinct type of item, such as a food, component, compositional value and source reference. Each entity contains a set of descriptors/properties, e.g. name of a food or component. The main entities and their properties are described in this text, with detailed specifications available in the Technical Annex.

The proposed EuroFIR recommendations should form the basis for a European standard that could be adopted within e.g. the CEN framework. It should be robust and flexible enough to incorporate future extensions with respect to components and other data. Standardised terminologies/thesauri should be established and used wherever these can increase the comparability of data.

The standard should ideally cover all steps in the data collection and compilation process, including description of sampling procedures, sample/food description, and analytical procedure or derivation method for compositional values, source of data and quality criteria for the resulting compositional value. Rules for calculation of recipes including factors for changes in weight and component levels during preparation are other features that should be considered.

2 Framework for food composition data

The standard should be applicable to all stages in the data compilation and management process, including the description of the sampling procedures, as well as data interchange. These and other steps in the process may involve the exchange of information on foods that does not include compositional data. However, this version of the standard addresses the requirements of food composition data, in which the presence of compositional values is considered mandatory.

The following terms will be used within the framework:

Data Management

In its broadest sense, data management includes all procedures relating to the handling of information, in this case relating to food composition, including offline and hardcopy

functions as well as the management of computerised data. Computerised data management covers all activities relating to the input, maintenance, processing and output of information relating to food composition, preferably handled within specialist software usually referred to as a Food DataBase Management System (FDBMS).

Data Interchange

Data interchange is the process of exchanging data between computer systems, which requires that the receiver as well as sender understands the data specifications in respect of format, structure and content. All data interchange conforming to this standard will use formats based on XML (see Technical Annex). In addition to the transfer of composition data, data interchange may involve other types of data, for example sets of bibliographic records, or the return of information from a thesaurus.

Interchange Package

An interchange package is a single valid set of interchange data that contains information defining its source and content. An interchange package may include external links, for example to thesauri defining terms and codes contained in the package.

A model for food composition data

The basic structure of food composition data is modelled on the metaphor of a food composition table (Figure 1). The upper left quadrant of the table represents information that describes the dataset as a whole, such as ownership and conditions of use. This information should be included in data interchange files, but may not be parts of the data that are actively handled within the database management system of the data compiler.

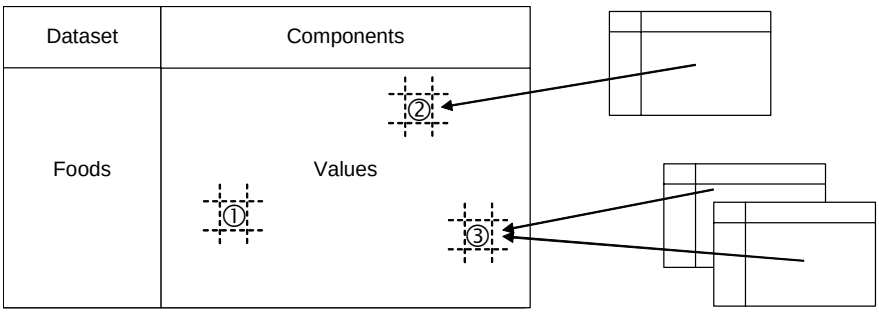

Fig. 1: The table metaphor (modified from Schlotke *et al.*, 2000)

The Food, Component and Value parts of the framework also hold additional descriptive information on these items. There are three basic sources for compositional values (also see Fig. 1):

1. A value may be an analytical, calculated, or estimated value original to the dataset.
2. A value might be a copy of another value from the same or another dataset.
3. A value might be an aggregate of multiple other values, taken from the same or another dataset.

For an original value, the method used to obtain it should be documented. When a value is obtained by copying or aggregating contributing values, appropriate method documentation for the contributing values should be recorded, as well as the reference to the original source of the contributing value(s)

In order to describe foods, their composition and their properties, it is necessary to define some basic categories for the information (entities). Each entity is characterised using a set of descriptors/properties, e.g. name of a food or a component. The following main entity categories are proposed to be included:

- Food
- Component
- Value
- Method specification
- Reference

In data management of the information relating to these main entities, the use of additional entities may be necessary. For example, the overall Food entity may include an entity that records food names separately, with each name linked to the associated food item.

In interchange files, the food composition data need to be prefaced with information on the Sender and Content of the file. The standard should include indicative documentation to be used in data interchange where the detail cannot be included, for example confidential recipe information.

Data levels

Based on data compilation guidelines (Greenfield & Southgate, 2003), food composition data are managed at different levels during the compilation process. A modified view is presented here, which also takes account of the model used by USDA (Haytowitz *et al.*, 2007). Four levels relate to the compilation of, for example, national data resources and a fifth relates to the adaptations and developments that major data users such as food consumption surveys or software providers might implement.

Level 1. Data sources contain published and unpublished research papers and laboratory reports containing analytical data, in the form presented by the original authors. The data might be systematically managed within a Laboratory Information Management System (LIMS), although in this case characteristics of level 2 data may be included. Food product information from manufacturers, including nutritional labelling data, is also included at this level.

Level 2. Initial databases contain original data from each data source. The data may be expressed as they were originally reported or converted into standard units. Information may be translated into standard coding or naming schemes, for example for food and component description, at input or during editing procedures that also involve further evaluation of the data. Data for individual analyses and food samples are held separately, so more than one value may be held for a component for a food item that will appear in the published database. Enough documentation should be held for each value so that it is unnecessary to refer back to the original sources, giving details of origin and number of food samples, food and analytical sample handling, edible portion, waste, analytical methods and quality-control methods.

Level 3. Aggregated and compiled databases contain data reporting a single value for each food/component combination, aggregated where appropriate from multiple values in the level 2 database. Links back to the food samples and component values used in the aggregation should be maintained, providing documentation on methods, sampling procedures, bibliographic references, laboratory of origin, date of insertion and other information relevant to the compilation process. The aggregated database includes all foods to be included in the

published database and all components, including those that contribute to derived component values reported in the published database but not necessarily included in it.

In addition to the level supporting the aggregated data, a further level defined as *Compiled* may be introduced, in which values are created within the system to complete missing data (e.g. logical zeros, calculated values). When data are needed for expert groups, one does not go back to the initial data and re-aggregated data every time: the compiler can refer to the aggregated/compiled database. There may be more data in the aggregated/compiled database than data published (e.g. individual sugars or fatty acids).

Level 4. Published databases and tables: the public resources which hold evaluated food composition data, normally disseminated by national compilers. Published databases and tables may reveal only a small part of the databank, being subsets or derivations of the aggregated or compiled database, and may be specially designed to meet the needs in terms of form and content of different user groups. They may include data that have been weighted or averaged to ensure that the values are representative of the foods in terms of the use intended. Data for extra foods may be added from raw-to-cooked or recipe calculation.

Level 5. User databases: adaptations of the published databases that major users such as food consumption surveys or software providers may make to tailor the data to their particular purposes, for example in the foods and components covered. They can also be data sets that the compiler has extracted from the databank for specific users (e.g. consumption surveys, software providers). If a completed matrix is required, any remaining missing values might be estimated, possibly to lower levels of certainty that would be included in the published database. Data might be derived for foods that more closely corresponding to the consumed foods, for example through further recipe calculations.

The level 2 and 3 databases are normally maintained through a computerised food composition database management system (FDBMS). A single database may support the Initial, Aggregated and Compiled data levels, with the FDBMS managing the separate levels. It is from this system that the published databases and tables are prepared. The eventual standard will cover all the requirements for the FDBMS, but the first versions will concentrate on the specifications necessary for the Aggregated and Compiled levels. Thus some areas such as food sample documentation will not be fully elaborated in these versions.

Standardised Terminology/Vocabulary

Standardised terminology is maintained as controlled language vocabularies listing the agreed or standardised terms. Such lists, commonly called thesauri (or controlled/standardized vocabularies), often include cross-references and scope notes. They may be organised in a hierarchy, in which case the terms are assigned *broader term* and *narrower term* relationships.

Each standardised vocabulary is generally maintained and published by some authoritative body. Examples are names of countries and languages, units, classifications (e.g. food groups), analytical methods (e.g. AOAC, NMKL), etc. Authoritative bodies include ISO, CEN, CODEX, INFOODS, EUROFOODS and the LanguaL Technical Committee.

In EuroFIR, thesauri for components, methods and thesauri supporting value documentation will be established and maintained [by a technical committee]. Rules for updating will be established. The official EuroFIR thesauri will use English as their main language. It is up to each user to translate thesauri for local usage. However, it is recommended to establish a

central authority within each country, or groups of countries with common languages, to maintain and publish translations. These should follow established rules for updating. EuroFIR should keep track of existing translations. This information should also be accessible on the Internet.

The proposed EuroFIR thesauri are:

- Components
- Units
- Matrix units (modes of expression)
- Value Types
- Method Types
- Method Indicators
- Publication Types
- Acquisition Types

The LanguaL thesaurus will be used for food description and classification. In addition, standards on language and country specifications will be used, and lists with component groups and component-method combinations will be developed.

Priorities for properties

The priority for a property (reported in XML as an attribute or as element content) defines whether the information is mandatory, recommended or optional. The priority will often depend on the database level, with more documentation required at the earlier levels, closer to the original source of the data. The priorities defined in the specifications in the Technical Annex to this document are based on the information required by food composition data compilers, i.e. that in a level 3 aggregated or compiled database.

There are two priorities:

1. *Mandatory* (M) properties are required as they build the core set of documented data that is needed to properly identify and describe the food composition data.
2. *Optional* (O) properties provide additional information describing the data and thereby assisting the user to assess them.

Priorities are also given for whole entity sets (i.e. database tables). If an *Optional* entity set is used, the priorities for its properties apply as indicated in that entity set.

A property may not have the same priority in all circumstances or contexts. For example for the REFERENCE entity, the property ISBN might be defined as *Mandatory* when the publication type indicates a book, but it is not relevant for a journal article. These variations are noted in the detailed specifications.

On occasions, *Mandatory* information may not be available, but in this case it should always be reported as such. Many properties have bespoke procedures for this, e.g. the use of "Anon." for missing authors. Thesauri that are used for properties should always include the term "Unknown". They may also include a term for "Other", but the use of this normally indicates that the thesaurus needs to be upgraded.

Data Type Formats

The Eurofoods recommendations (Schlotke *et al.*, 2000) specified data types in terms of the requirements for delimited and fixed-length text files. The constraints of an XML format differ and revised definitions of the data types available in the XML interchange files are given in the Technical Annex. These definitions specify the information that can be incorporated into an interchange file and individual designers of FCDMS must decide how compatible data should be managed internally.

3 Definition and description of main entities

The main entities in a set of aggregated or compiled food composition data are shown in Fig. 2 and described in the following sections.

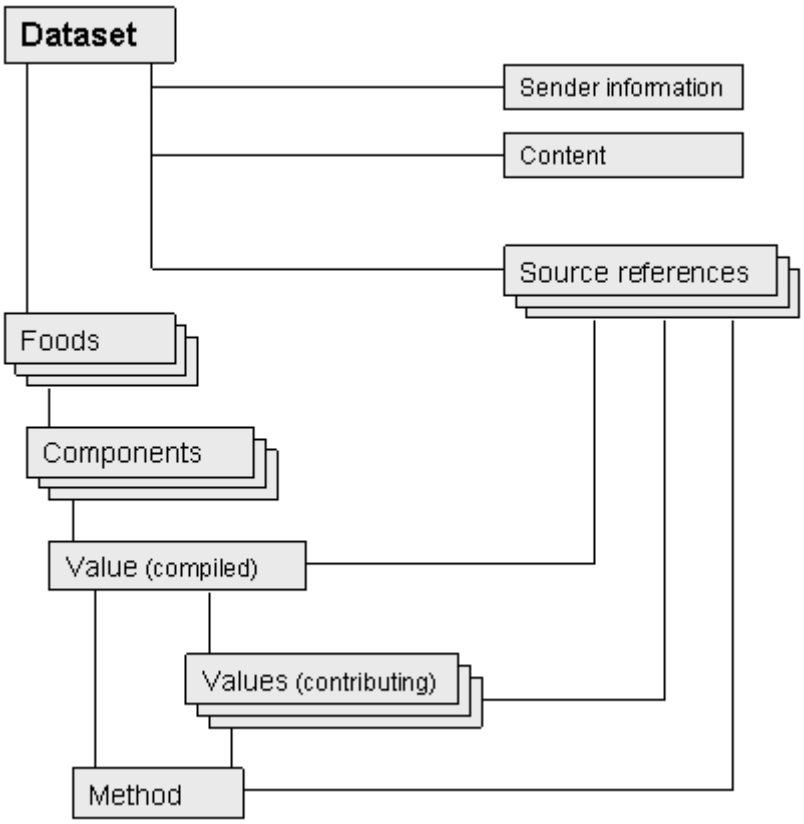

Fig. 2: Schematic chart of main entities and their relations in an aggregated or compiled food composition dataset

FOOD entity

The FOOD entity is *Mandatory* and contains properties that describe the foods reported in the database. A large set of descriptors and supporting information is necessary to properly identify and describe a food in a database, such as its name, classification and description according to standardised thesauri (e.g. LanguaL), see Technical Annex. In addition information specific to particular samples analysed in a laboratory, food products from a specific producer, as well as to more generic foods and products may be recorded. The EuroFIR standard supports the retrieval of corresponding food items across conforming databases, which requires the use of a standardised terminology to describe the food items.

Thus a main element of food description will be provided by LanguaL indexing (http://www.langual.org). Some types of food item, for example food sample records, may require additional description.

There is no universal agreement on food classification as different systems serve different purposes, hence it will be necessary to map the different classification systems currently used in European databases. A EuroFIR classification system will be established for use in food composition data interchange. International standard classifications are incorporated into LanguaL facet A, so terms can be assigned as part of the LanguaL indexing process.

The Eurofoods recommendations include an optional supporting entity for Contributing Foods that can be used to link a derived or aggregated food item to all its contributing foods and their description. It can also support a simple ingredient list. The EuroFIR standard should ideally include facilities for detailed listing of the ingredients for a recipe, data on yields (fat/water) and retention factors for components during preparation. Facilities to document rules for recipe calculation should also be considered.

COMPONENT entity

The COMPONENT entity is *Mandatory* and reports food constituents that can be measured by chemical, physical and microbiological methods or can be calculated from the measured constituents, and include nutrients, bioactive substances, and contaminants. Other measures and properties such as density, per cent edible portion, or pH may conceptually also be handled as components. Food specific factors to be used for calculated constituents may also be modelled as components (e.g. nitrogen conversion factors for protein calculation), although strictly these form part of the metadata for the resultant compositional value. The COMPONENT entity includes properties such as name (possibly in both the national language and English) and standard classifications.

A thesaurus with definitions and description of components will be included in the proposed EuroFIR standard. The component thesaurus will primarily cover nutrients and related bioactive compounds with possible positive health effects. There are several systems for identifying components. The EuroFIR thesaurus will be based on the Eurofoods component list with links to the ChEBI database and including other vocabulary such as INFOODS Tagnames. A hierarchical approach will, where appropriate, be used, from broad to more specific definitions of individual and derived components, e.g. fatty acid 18:2 < 18:2 n-6 < cis,cis 18:2, n-6. Derived components such as “saturated fatty acids” may differ among databases depending on which fatty acids are included. Different analytical or calculation methods may produce different values and may therefore have to be treated as reporting separate components.

VALUE entity

The VALUE entity reports a result for the amount of a COMPONENT in a FOOD, together with description of the value, such as its statistical properties. The amount may have been obtained by chemical analysis, calculation or imputation (estimation). The method used will influence the value description that needs to be reported, as well as the related method and reference documentation.

The VALUE entity is *Mandatory* in a food composition dataset and contains attributes that describe the value reported for a food-component combination. These include the value type (e.g. “mean”, “less than”, “below detection limit”), derivation method and quality rating index. Statistical information, such as the number of samples, mean, median, minimum,

maximum and, where appropriate, standard deviation may also be included for analytical data, as described in more detail in the Technical Annex. The VALUE entity is linked to entities reporting method and reference documentation for the value, as depicted in Fig. 2 and described below.

An attribute called “selected value” is included to store a single figure as the best representation of the statistics, based on the decision of a data compiler. The numerical amount should be stated to the correct number of significant figures including trailing zeroes, for example a typical value for Nitrogen might be '3.90' and this must be reported as such, not just as '3.9'. Extra decimal places obtained from analysis or calculation may be retained internally, but should be appropriately rounded when the value is interchanged or published. Rounding within integer values may be applied editorially by the compiler. Consistent policies should be applied where the figure reported might depend on the units used, for example a carotene value expressed as 1.03 milligrams may be a modification of 1031 micrograms, as usually expressed.

Some metadata proposed in Schlotke *et al.* for the COMPONENT entity are now linked to the value as they relate to the way the numerical value is expressed. These are the *unit* (mg, g) and *matrix unit* (previously called the *mode of expression*), which refers to the measure of the matrix within which the reported component occurs, e.g. *per 100 g edible portion* or *per 100 g total fatty acids*. Although these metadata refer to individual values, there remains the option of declaring a default unit for a component's values throughout a dataset. A default matrix unit is more difficult, because in some cases a fixed matrix unit will be associated with the values for a given food, whereas for some analytical values a particular matrix unit will apply to given components.

The VALUE entity also includes basic method documentation for the value, including Method Type and Method Indicator terms. Optionally, more detailed information can be provided for analytical values through the METHOD SPECIFICATION entity.

The Eurofoods recommendations did not give any proposal for a system to assess data quality, but these will be included in the EuroFIR standard.

METHOD SPECIFICATION entity

The METHOD SPECIFICATION entity reports a detailed specification of the analytical method used to obtain the value that links to it. In addition to properties for official method and method reference information, it includes sections reporting the analytical key steps and laboratory performance parameters. The METHOD SPECIFICATION entity is *Optional*.

REFERENCE entity

Bibliographical and related information on the original references cited in the dataset are recorded through the REFERENCE entity. References from the published and unpublished literature (e.g. laboratory reports) may be associated with component values, methods, food/sample description and other metadata. A REFERENCE item must be described with sufficient bibliographic reference information in order to be uniquely identified, although this may be difficult for some grey literature such as food labels.

The REFERENCE entity holds information on references cited in the dataset for entities such as values and methods, as well as references cited within Remarks and other textual fields[1]. REFERENCE items are normally linked with entities that cite them either through a database link table (if a many-to-many relationship is to be supported) or through a field in the entity table if the entity only requires a single reference to be cited. A VALUE may link to multiple REFERENCE items if the result is a value averaged from several sources. On the other hand, a link field may be sufficient for FOOD (e.g. reporting a food sample or recipe) or for METHOD SPECIFICATION records.

A reference may appear in one of various bibliographic forms such as published or unpublished reports, journal papers, articles in books, web-pages, food labels, etc. This aspect is coded using the Reference Type thesaurus. A further attribute, Acquisition Type, is used to record the status of the information reported, such as peer-reviewed scientific publication, evaluated food table data, 'own' data or food manufacturer data.

REFERENCE is a *Mandatory* entity (assuming that at least one reference is cited in the dataset). The fields should correspond as far as possible to those conventions generally used in the scientific literature (e.g. Vancouver system) and included in standard software for managing and publishing bibliographies (e.g. Reference Manager or EndNote).

SENDER INFORMATION

In data interchange a header element, SENDERINFORMATION, of a file reports on the person and/or organisation that prepared and despatched the Food Transport Package and information about how to contact them.

CONTENT

In data interchange a header element, CONTENT, of a file reports element information about the food data contained in the FOODS element of the Food Transport Package. It provides the name of the dataset represented by the content of the FOODS element and of the organisation or body responsible for it. It specifies any legal restrictions, e.g. copyright, concerning the data and summarises the content of the package. It also provides the overall bibliographic reference to the source of the data held in the FOODS element.

4 References and links

Charrondière U.R., Vignat J., Møller A., Ireland J., Becker W., Church S., Farran A., Holden J., Klemm C., Linardou A., Mueller D., Salvini S., Serra-Majem L., Skeie G., van Staveren W., Unwin I., Westenbrink S., Slimani N., Riboli E. (2002) The European Nutrient Database (ENDB) for Nutritional Epidemiology. *J. Food Compos. Anal.*, **15**, 435-451.

Greenfield, H., Southgate, D.A.T. (2003). Food Composition Data: Production, Management and Use. 2nd Edition, FAO, Rome. Available from http://www.fao.org/infoods/publications_en.stm.

Gry J., Holm S., Morgan M., Møller A., Preece R., Speijers G., Søborg I. (2002). EU-FAIR Concerted Action CT 98-4419. BioActive Substances in Food Plants Information

[1] This requires that a reference identifier is incorporated into the text as the link to the associated REFERENCE record, for example marked up as <RefCode>A0001</RefCode>.

System (BASIS). Final Report (1999-2001) Scientific Groups' Report. Danish Veterinary and Food Administration, Søborg, Denmark. See also http://www.foodcomp.dk/basis/ .

Haytowitz D.B. et al. (2007). To be filled in when citation becomes available after IFDC, October 2007.

ISO (2005). ISO 22000:2005. Food safety management systems -- Requirements for any organization in the food chain. Available at http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=35466&ICS1=67&ICS2=20&ICS3=

Schlotke F., Becker W., Ireland J., Møller A., Ovaskainen M.-L., Monspart J., Unwin I. (2000). *EUROFOODS Recommendations for Food Composition Database Management and Data Interchange*. COST Action 99 – EUROFOODS Research Action on Food Consumption and Composition Data. COST Action 99 Report EUR19538, European Commission, Luxembourg. See also: *J. Food Compos. Anal.*, **13**, 709-744.

5 Organisation and project acronyms

Acronym	Name	Web address
AOAC	Originally *Association of Official Agricultural Chemists* and then *Association of Official Analytical Chemists*, now the legal name is AOAC International	www.aoac.org
BASIS	BioActive Substances in Food Information System	http://www.polytec.dk/ebasis/
CEN	European Committee for Standardization	www.cenorm.be
CODEX	Codex Alimentarius	www.codexalimentarius.net
EPIC	European Prospective Investigation into Cancer and Nutrition	http://www.iarc.fr/epic/
ENDB	EPIC Nutrient DataBank	
FAO	Food and Agriculture Organization of the United Nations	http://www.fao.org/
FDA	U.S. Food and Drug Administration	http://www.fda.gov/
ISO	International Organization for Standardization	www.iso.org
NMKL	Nordic Committee on Food Analysis	www.nmkl.org

Acronym	Name	Web address
USDA	U.S. Department of Agriculture	www.usda.gov