ELSEVIER

PII: S0308-8146(96)00153-7

Food Chemistry, Vol. 57, No. 1, pp. 149–154, 1996
Copyright © 1996 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0308-8146/96 $15.00+0.00

The Component Aspect Identifier for compositional values

Ian D. Unwin[a]* & Wulf Becker[b]

[a]*Food Information Consultancy, The Opas Centre, St John's Innovation Centre, Cowley Road, Cambridge CB4 4WS, UK*

[b]*National Food Administration, Box 622, 751 26 Uppsala, Sweden*

The Component Aspect Identifier (CAId) concept was developed to define the key information relating to compositional values for foods and to provide convenient mechanisms for handling this information in a Food Database Management System (FDBMS). The CAId identifies the component, defines the basis on which the value is expressed (Mode of Expression) and summarizes supporting information on the origin, method, source and quality assessment for the value. The Method Indicator of the CAId summarizes the analytical method for analysed values or the basis of the calculation for derived values; a notation is proposed for organizing the analytical method information. The CAId has been implemented in the EuroNIMS food information management system which handles analytical, calculated and imputed values, as well as a wide range of information on food items. The CAId is discussed in terms of the level of specificity represented in information describing foods and reporting their composition. Copyright © 1996 Elsevier Science Ltd

INTRODUCTION

A value for the amount of a component in a food needs to be associated with information on the food and the value itself. Various methods for the categorization, description and naming of foods have been described (e.g. Kohlmeier, 1992; Truswell *et al.*, 1991; Unwin, 1992). Food description using LanguaL (Hendricks, 1992) is probably the most systematic of these and uses a faceted approach in which particular aspects are indexed by descriptors selected from different hierarchies, for example for food source and for processing applied to the food.

The description of the INFOODS Data Interchange Format (Klensin, 1992) specifies a data structure which incorporates component identification for a value. The identifier includes aspects of analytical method, calculation method and the basis on which a value is expressed, in those cases where these factors affect the compatibility of the numeric values quoted.

Data structures recording similar information about values within a Food Database Management System (FDBMS) have also been reported (Feinberg *et al.*, 1992). The NUTSYS project of the Swedish National Food Administration (Becker, 1993) developed and validated a model for a relational FDBMS operating with Structured Query Language (SQL). The Swedish NFA contributed the NUTSYS results to the collaborative project to develop the EuroNIMS Food Information Management System and initiated further work to investigate the requirements for handling information on the component aspects of a compositional value within an FDBMS (Unwin & Becker, 1995). The present paper describes the main proposal resulting from this work, the Component Aspect Identifier (CAId) and its implications for the data structures and facilities which may be implemented in an FDBMS.

INFORMATION HANDLING

In data interchange it is essential to differentiate component values which are, or may be, incompatible. In the INFOODS format this is done by assigning different tagnames when analytical methods measure different chemical entities, for example for 'fibre', or when values are expressed differently, for example carbohydrates as weight or as monosaccharide equivalents. Handling compositional data within an FDBMS require greater flexibility to enable values expressed differently to be displayed in a comparable form and for details of analytical method to be held separately from the component identification, even though this and the method may be correlated. FDBMS facilities should allow the

*To whom correspondence should be addressed at: 6 Chapmans Way, Over, Cambridge CB4 5PZ, UK.

database compiler to review whether values can be considered comparable or in some way represent non-equivalent measures.

Within a relational DBMS, data are structured into separate 'tables' linked through common values in 'key' fields which identify specific records. Thus, component value records are usually linked to records in two further tables, one containing records for individual foods and the other for individual components. Further tables might be linked to a component value, for example with details of the analytical method or bibliographic reference. Generally when linked records are retrieved in response to a query, a temporary table is generated which, within its rows, combines relevant fields from the underlying tables. The user may be permitted to edit fields in this temporary view of the data (displayed as the table or as single formatted records), the changes being stored in the underlying tables.

This is a useful mechanism for viewing data stored in relational structures when the user is simultaneously interested in a set of fields taken from several tables, but does not routinely need to see further fields. However, for the records associated with a component value the user may not be interested in, or may be familiar with, the detailed fields for one aspect of the linked information or will need to view the full linked record when evaluating and comparing values. Since reviewing lists of values is an important function to be supported by an FDBMS, a method is essential for providing the user with a convenient summary of, and access to, the data associated with each value. This requires an organization and representation of the underlying data in a summarizing structure that is meaningful to the user and presented through a well-designed interface. This structure may use stored data or data derived from them.

Organizing data review procedures around a summarizing structure also provides benefits in system design and system operation. It defines routes by which data will be accessed so that parameters can be established, for example that an accessed record should be read-only. It provides an interface to data that can be user-defined, for example alternative ways of identifying a bibliographic citation might be made available since it is unlikely that the record identifier itself will be meaningful to the user.

Information associated with compositional values identifies the component, defines the basis on which the value is expressed and provides supporting information on the origin, method and source reference for the value. The present work distinguishes the following categories for component value information:

- component identity;
- mode of expression;
- origin of value;
- method of generating value;
- source of value;
- source reference;
- quality of value.

COMPONENT ASPECT IDENTIFIER

The above categories of information associated with compositional values have been combined into a single representation, the Component Aspect Identifier (CAId). Each part of the CAId displays information on one of the categories. Several parts use identifiers pointing to more complete details in underlying data such as literature reference records.

Parts of the CAId, in particular the Component Identifier, the Mode of Expression and the codes used to indicate calculation method, are based on the INFOODS tagname system (Klensin *et al.*, 1989) used for identifying food components in data interchange. However, such aspects are treated as independent entities in the CAId, although their representations can be implemented to maintain as much compatibility and interconvertability with the INFOODS tagnames as possible. The CAId holds more information (for example on those analytical methods which are considered equivalent when assigning tagnames) than the INFOODS system. As a result it should be possible to maintain a unique mapping from CAId data to INFOODS tagnames but the reverse translation may require information from other fields in the INFOODS Interchange Format.

Each of the categories of component-related information, together with the way it is implemented in the CAId, is described below.

Component Identifier

This part of the CAId identifies the component independently of value-related aspects such as mode of expression and analytical method.

The purpose is to collect together all information relating to the same component, and to distinguish between values measuring different components. For example, any measure of total quantity for a vitamin, whether obtained by bioassay or the summation of individual vitamer activities, would be associated with a single Component Identifier. On the other hand, the individual vitamers are considered separate components, including a predominant one to which the overall activity is related (as 'equivalents'). Vitamin A, by bioassay or the summation of vitamer retinol equivalents, has one Component Identifier and retinol itself a different one.

The availability of generalized FDBMSs such as EuroNIMS will increase the number and types of food components for which values are stored. Data on naturally occurring non-nutrients, additives and contaminants will require a more sophisticated identification of components than is often used at present; even for nutrients this is already a problem, for example with fatty acid isomers. The underlying component identification used for stored values should be unique and unambiguous (where it can be made so, for example by basing it on a suitable representation such as chemical structure) within the system and capable of correct translation during data import, data export and in interactions with the user. Use of the CAId provides flexibility in the choice of Component Identifier displayed to the user to represent the underlying form. For example, component identification using Chemical Abstracts Service Registry Numbers might be displayed as meaningful abbreviated component names.

Component Identifier codes in their initial EuroNIMS implementation, EuroNIMS version 0.9, have been based as far as possible on the existing INFOODS tagnames, where necessary truncating these before the part that contains information on mode of expression or analytical method since this is represented in other parts of the CAId. Comprehensive FDBMS facilities are considered to include the capability of storing values when these are reported, so a Component Identifier must be assigned when a component is first encountered. The solution to this and the problem noted above of covering an increasing range of compounds may be to include a structure-based Component Identifier within an FDBMS, maintaining a mapping to registered tagnames and possibly using tagnames for substances not well defined structurally.

Mode of Expression

This aspect refers to the alternatives which may be used to express the amount of the component present in a food. The normal Mode of Expression is unit weight of component per unit weight of food but others include expression:

- on an alternative basis to component weight, e.g. carbohydrates as monosaccharide equivalents;
- in terms of other components, e.g. amino acids per g nitrogen;
- in terms of alternative food measures, e.g. 'per unit dry weight' or 'per unit volume';
- as percentage, e.g. energy, percent contributed by protein;
- as a ratio of components, e.g. polyunsaturates to saturates.

Alternative units, for example values given in g rather than mg, are (arbitrarily) not considered different modes of expression.

In the EuroNIMS 0.9 implementation of the CAId, the Mode of Expression identifiers are single uppercase alphabetic characters, for example M for values as monosaccharide equivalents and N for amino acids per g nitrogen. It is possible for a value to require identifiers from more than one of the categories listed above; if this is the case, ordering rules are applied to standardize the notation.

In general, values representing different modes of expression for a component can be interconverted, although this may require a value for a related component (e.g. for nitrogen where amino acids are expressed per g nitrogen). The Mode of Expression identifiers will control the FDBMS facilities for interconverting values. The system may convert to a standard mode of expression, making them transparent in the displayed values but then requiring them to be reported for the underlying stored values.

Expression on an alternative basis to component weight may be a more general situation than is catered for in this version of the CAId, in such cases as β-tocopherol expressed as α-tocopherol equivalents. Some modes of expression involve more than one component. A ratio mode requires two Component Identifiers. A second Component Identifier is also implicit in values expressed in terms of other components; information on the units of the second component becomes significant in this situation. Mode of Expression identification will need further development to accommodate these cases.

The Origin and Source Type Flags

These two important elements of the CAId appear similar but the difference between them is crucial. The origin of a value is the means by which it was generated (analysis, calculation or imputation) and is associated with a method. On the other hand, the value source is where the user or system obtained the value and is associated with a source reference.

The Origin Type Flag and the Source Type Flag are codes (consisting of single lowercase alphabetics in the initial EuroNIMS implementation), which categorize the origin and source of a value. For instance, the origin of a value might be recipe calculation with its source being importation from another EuroNIMS system.

Each of the two flags is closely related to further parts of the CAId dealing with the origin and source of the value. The Origin Type Flag information is expanded by the Method Indicator and the Method Pointer. The Source Identifier links to source reference information appropriate to the Source Type flagged for the value.

Method Indicator and Method Pointer

The method information associated with the Origin Type Flag consists of two elements, the Method Indicator and the Method Pointer. The Method Indicator is a meaningful keyword representing the method by which the value was obtained while the Method Pointer registers the link to detailed method records.

For analytical values, the Method Indicator records the general analytical method used. It holds the 'headline' method (a term classifying the main feature of an accepted method for a component, e.g. 'bioassay', 'HPLC'), with the option to append qualifying terms indicating, for example, the organism or method of detection, building a notation as described in more detail below. The Method Indicator summarizes (or indexes) the method, but is separate from the detailed documentation on analytical method held in the FDBMS. The Method Pointer is available for linking to these records. For example, in EuroNIMS version 0.9 it points to a source reference for the analytical method.

	Component	Mode of	Origin			Source		Quality
	Identifier	Expression	Flag	Method		Flag	Identifier	Index
				Indicator	Identifier			
Acronym	*COMP*	*MODE*	*OTF*	*MethInd*	*MethID*	*STF*	*RefID*	*QInd*
Example	STARCH	M	a	Polarimetry	J0002	f	B0007	-

Fig. 1. This shows the structure of the CAId and gives an example of its content. The example CAId should read as 'Starch expressed as monosaccharide equivalents: analytical value from food table, reference B0007; determined by polarimetry as in journal article J0002'. Thus the CAId summarizes the key information documenting a compositional value and can be displayed as a compact character string with the parts separated by delimiting punctuation characters.

For calculated component values, the Method Indicator records the computation method used, for example the set of factors applied in calculating energy values. Codes can be used where appropriate, for example 'KJA' (kilojoule conversion factors using available carbohydrate); these may be based on those used in the INFOODS guidelines (Klensin, 1992). The Method Pointer can be used to link to a source reference for details of the computation, for example published factors.

Source Identifier

This part of the CAId identifies information on the source reference for a component value. In EuroNIMS 0.9 the identifier for the reference reporting the value is used (whereas the Method Pointer is the identifier for the reference reporting the method, although the two might be the same). Future enhancements may provide the capability of substituting a meaningful document identifier, for example 'Author (year)', for display of the Source Identifier.

Quality Index

This part will provide an indication of data quality for a value. For analytical values it will be a Quality Index evaluating criteria such as method, laboratory and information on sampling. For derived data it may be used, for example, to indicate that a component value from recipe calculation included an ingredient with a missing value. Further investigation is needed to specify the quality index requirement in more detail.

CAId display formats

In practice most FDBMS implementations would use an enhanced, graphical formatting of CAId information, providing 'hot' links to underlying data where appropriate. In EuroNIMS version 0.9 clicking on items such as the Source Identifier opens the underlying record, in this case that for the source reference.

ANALYTICAL METHOD NOTATION

A convenient way of summarizing analytical methods was required for the CAId. Results from food composition analysis are often described by phrases such as 'determined by an HPLC method' or 'determined using bioassay', suggesting that these representations of method might encapsulate it in a 'headline' description. Thus, it was decided that the main part of a Method Indicator should be a single word, abbreviation, acronym or abbreviated phrase that would be widely recognized as the 'headline' method.

It was also recognized that it might be necessary to qualify this headline method with further information, either giving extra detail on that method or supplying information on an associated procedure considered important. An example of the first case might be 'HPLC, reverse phase' and of the second 'HPLC, fluorimetry', meaning 'HPLC followed by fluorimetric detection'. The appended phrase is referred to as the headline modifier.

In the EuroNIMS 0.9 implementation, the Method Indicator text was limited to a length of 22 characters. A sample set of analytical method records was created from methods reported in food tables and Method Indicators were assigned. The length constraint required that every option for shortening was applied, particularly if a headline modifier was needed. For these, a concise way of indicating the relationship between the headline method and modifier improved the information content. Further, for consistency and to aid user recognition of shortened forms, it was considered that these should be used even where the shortening was not necessary to meet the maximum length requirement. As a result, the Method Indicator text became a list of short terms representing analytical procedures, the most important listed first followed by modifiers, each of which was preceded by an indication of its sequential relationship to the headline method.

The relational punctuation is used to indicate these relationships between the parts of the Method Indicator text. Since the headline method at the beginning is the most important (or most definitive) aspect of the method procedures, not necessarily the first of the analytical procedures recorded for the overall method, the character ' < ' is used to indicate 'preceded by' just as ' > ' indicates 'followed by'. Relational punctuation currently defined for Method Indicators is shown in Table 1. The headline procedure is terminated by the first occurrence of ' < ', ' > ' or ':' relational punctuation. The various items contributing to the headline modifier

Table 1. Relational punctuation in analytical method notation

>	introduces the next procedure or indicates a later procedure omitted	HPLC > fluorim
<	introduces a procedure earlier than the headline method or indicates an earlier procedure omitted	GLC < derivn
> >	introduces a later procedure, with intervening procedures omitted	oxidn > > fluorim
< <	introduces an earlier procedure, with previous procedures omitted	fluorim < < oxidn
< >	introduces a later procedure, with ones before the headline method omitted	hydrol < > colorim
:	introduces a simultaneous procedure or the subject of the procedure	GLC:acetate
,	introduces a term qualifying the procedure, making it more specific	HPLC, rev ph
()	encloses a name of a modified or more specific form of the procedure	gravim(AOAC)
[]	encloses a synonym more familiar than the preferred term	
{}	encloses a code such as an INFOODS keyword	{CNT}

follow in strict chronological order. Other punctuation introduces a modifier of the preceding procedure.

The Method Indicator provides flexible possibilities for the matching and retrieval of methods. The procedures can be reorganized into full chronological order, providing a form independent of the procedure selected as the 'headline method'. Matching can include or ignore the presence or absence of individual procedures (whether these were not used in the analysis or simply not recorded). Used in conjunction with a hierarchical thesaurus of procedure terms, matching is possible even for procedures recorded at different levels of specificity or as synonyms. Results could be listed in order of nearness of matching. A Quality Index could be assigned on the basis of the level of detail provided. Initial experience with the use of Method Indicators in EuroNIMS 0.9 may indicate which of these possibilities will be practical and useful developments, although it may be necessary to provide separate software for handling the notation and thus to test it thoroughly.

DISCUSSION AND CONCLUSION

The CAId is designed to organize and concisely summarize key information handled by an FDBMS for component values. The data for the various parts of the CAId may be stored in the form displayed or may be derived from underlying related information associated with the value. Thus, there is the potential to modify the displayed data, perhaps under user control. For example, the user might be able to define the content and format of a meaningful document identifier, as noted in the section on the Source Identifier.

This paper has presented the organization of the CAId rather than detailing the relationship of its content to the structure of the underlying data storage. The CAId aims to bring together into a structured descriptor the information about a component value which a user needs at a level that is useful in comparing that value with other values. Although content has been defined for an initial implementation in EuroNIMS, the method of handling information for each aspect of the CAId, and more so the detailed CAId values, will develop as experience is gained from practical use of the system. Indeed the CAId provides a logical place in an FDBMS to translate a representation of information from that stored to that most meaningful to the user. Content may be defined in terms of translation lists or algorithms between the two, with the further option to apply different translations specific to a group of users, individual users or particular tasks being performed by the user.

Perhaps one of the more interesting features of the CAId is that different types of information are used in its various parts. There are simple categorizations implemented as flags, specific indexing notations as in the Method Indicator and unique identifiers for components and source references. These represent links to specific records in the underlying relational data structure. Equally, precise data could be accommodated if deemed appropriate, for example the method information could include number of samples and indeed the value itself might be presented as part of the CAId.

However, this heterogeneity is more apparent than real. Each aspect is potentially a hierarchy from less to more specific representations of its information; the form which the CAId representation takes depends on the level of detail which will provide the user with the most useful summary when comparing values. Indeed main sections of the CAId may display several levels in their various subsections. For example, the original of a value is represented at the levels of Origin Type Flag, Method Indicator indexing and the method record identifier (or source reference identifier). This is useful because the first level sets the context, indicating both to the system and the user what form the more specific information will take. The indexing level summarizes the information, providing meaningful content on which to retrieve, sort and compare records. The level of detailed information identifies a specific record or, at a more specific level, reports data from individual fields. For some aspects one level may be sufficient whereas in others it is necessary or helpful to define additional ones.

The CAId was developed to organize information associated with values reported for the composition of foods but the approach may be applicable to other types of information. This is likely to be useful where the user needs to list and compare records which may vary in various aspects of their data, and to select those records which are acceptable within the judgement criteria being applied. For food-related information this may apply to the comparison of food sample detail, the aggregation of food items and, possibly, to the review of detailed analytical descriptions. The analytical method notation proposed in this paper may provide an indexing basis in this latter area.

Systems for describing foods, as referenced in the Introduction, often use a faceted approach. In particular, LanguaL consists of separate hierarchical facets (called factors) which index various aspects of the description. Application of a CAId-type approach to food description using LanguaL as its indexing language might permit more flexible solutions to the handling of information on food items while maintaining the integrity of a standard LanguaL language, for example by associating indexing of an item with precise data for it.

In representing the most significant information associated with a component value, the CAId is an aid both to the data management system in providing criteria on which to base its data manipulation procedures and to the user in selecting and reviewing values. The CAId can provide the basis for a more user-oriented data model than one considered only in terms of underlying data structures. Its make-up and content can evolve to reflect the most important user requirements for reviewing value records and, in doing so, should help define the key requirements for the system handling of information relating to component values. The approach may be applicable to other types of data handled within an FDBMS.

ACKNOWLEDGEMENTS

The authors are grateful to the Swedish National Food Administration for supporting this work and would also like to thank Dr Jayne Ireland-Ripert and Dr Anders Møller for valuable comments made on the proposals for the Component Aspect Identifier.

REFERENCES

Becker, W. (1993). NUTSYS—a food and nutrition composition and information management system. Report No. 11. National Food Administration, Uppsala, Sweden.

Feinberg, M., Ireland-Ripert, J. & Favier, J.-C. (1992). Validated data banks on food composition: concepts for modeling information. *World Rev. Nutr. Diet.*, **68**, 49–93.

Hendricks, T. C. (1992). LanguaL—an automated method for describing, capturing and retrieving data about food. *World Rev. Nutr. Diet.*, **68**, 94–103.

Klensin J. C. (1992). *INFOODS Food Composition Data Interchange Handbook*. United Nations University, Tokyo.

Klensin J. C., Feskanitch, D., Lin, V., Truswell, A. S. & Southgate, D. A. T. (1989). *Identification of Food Components for INFOODS Data Interchange*. United Nations University, Tokyo.

Kohlmeier, L. (1992). The Eurocode 2 food coding system. *Eur. J. Clin. Nutr.*, **46**(Suppl. 5), 25–34.

Truswell, A. S., Bateson, D. J., Madafiglio, K. C., Pennington, J. A. T., Rand, W. M. & Klensin, J. C. (1991). INFOODS guidelines for describing foods: a systematic approach to describing foods to facilitate international exchange of food composition data. *J. Food Compos. Anal.*, **4**, 18–38.

Unwin, I. D. & Becker, W. (1995). The Component Aspect Identifier. A tool for handling food component information in a food database management system. Report No. 9. National Food Administration, Uppsala, Sweden.

Unwin, I. D. (1992). Food naming and description using faceted descriptors. In: *Report of the Second Annual FLAIR Eurofoods-Enfant Project Meeting*, eds J. Castenmiller & C. E. West. FLAIR Eurofoods-Enfant Project, Wageningen, pp. 77–84.