CODATA Task Group on a Systematic Nomenclature for Foods in Numeric Databanks. (Derek D. Singer, Visiting Expert, National Cancer Institute)

1.1. Introduction

1.1.1. Purpose of Task Group

CODATA (Committee on Data for Science and Technology) is a Committee set up in 1966 by the International Council of Scientific Unions with the objective of improving 'the quality, reliability, processing, management, and accessibility of data of importance in all scientific fields covered by ICSU.'

The Constitution of CODATA provides for Task Groups to examine specific scientific topics.

1.1.2. Membership of the Task Group

It has been the aim in selecting members to include those who have a particular knowledge, expertise and experience in the nature of food and food technology or information systems or computing. Many of those invited to be members of this Task Group can lay claim to two or three of these qualities. Although CODATA Task Group members are not national or organisational representatives, but selected because of their knowledge and experience, it has also been the aim to cover as wide an area of the world as possible. Thus we have members from N. America, Western and Eastern Europe, Asia and the Far East. Nevertheless, we have none from the South Americas or Africa. To some extent this deficiency may be compensated by enlisting corresponding members and on other occasions, by extending special invitations and convening special meetings.

However, although membership is preferably restricted to about the number invited, proposals for other permanent members would be welcomed.

1.1.3. Meetings

Unfortunately, CODATA like many international organisations, has to consider costs, and the first attempt to establish this Task Group failed for because of difficulties of financing. However following a number of meetings of workers in this area convened by the National Cancer Institute, National Institutes of Health, USA, it became clear that current work at the Institute required collaboration and effort via a similar international committee. The proposal to CODATA was therefore resubmitted with the suggestion that NCI could finance some meetings coincident with other meetings convened as part of its own research program.

However NCI would welcome support from any organisation in any other country in hosting and supporting further meetings.

intérêt
financier
à aider l
cΔ

It should also be made clear that NCI will exchange data and software - within legal constraints - with any organisation in any country in order to further development of the language and its applications, as indeed, it has done in the past.

1.1.4. Relations with other Organisations

There are a number of ICSU (International Committee of Scientific Unions) who have been or will be informed of the establishment of the Working Group. These include IUNS (International Union of Nutrional Science), IUFOST (International Union of Food Science & Technology) and IUPAC (International Union of Pure and Applied Chemistry) and IUBS (International Union of Biological Sciences). There are also international information services who have an interest in this general area. These include IFIS (International Food Information Service) and the CAB (Commonwealth Agricultural Bureau) who are responsible for the production of Food Science and Technology Abstracts and Nutritional Abstracts respectively. Information on similar bodies is welcomed since it intended to inform them of the progress of our work and if necessary, enlist their aid.
One of the best attempts yet (although one still inadequate in many respects) to design a system for describing foods (see below) has been made by the USA Food and Drugs Administration (FDA) collaboration with the FDA's information scientists is likely. This is especially necessary because the FDA is also attempting to promote similar international collaboration - although their effort was conceived some two years after the first approaches to CODATA to set up this Task Group. Some members will in fact be discussing future collaboration with the FDA subsequent to the CODATA meeting subject to the decisions of the Group.

1.2. The Need for an International Food Language

1.2.1. The Problem

Over the past one hundred and fifty years or so, the sciences of food, nutrition and medicine have become numerate. It might be said that to qualify as a science, any area of intellectual endeavour must be so. J J Thompson (Lord Kelvin), the nineteenth/twentieth century physicist said that whenever he perceived any phenomena he felt compelled to measure it. However measurements are of little value unless the phenomena or materials to which these result apply are precisely described. Such descriptions can be made with varying degrees of ease. A statement on the wavelengths of the visible sodium spectra needs little in the way of qualification in describing the nature of the material, sodium.

pb voca

surmonté par Darc (5)

mais numeriq et biblio

Measurements of the absorption spectra of organic compounds present greater difficulties. Not only are there several different ways of naming compounds semantically, but none of these are satisfactory for use in computers for complex searches by classes of compounds or sub-groups. The use of WLN (Wissweisser Line notation) or Professor Dubois' DARC system has now overcome that problem. Sub-structure searching is common and sophisticated software exists for retrieval of information on chemical substances and graphics display of their structures. Numeric databases can be linked to bibliographic data bases for retrieval for aditional information.

∃a de X & accordées en fct des désirs
pb: local et spécif
⇒ incomplet

In recent years, an increasing amount of numerical data on food has been determined and assembled in databanks for subsequent retrieval according to the desires of users of that data. However such numeric systems are local and tend to be specifically designed for the purpose of an individual or a single organisation or locality where the nature of the food is better understood. However, even in these circumstances many characteristics of the food are not recorded.

pb →

enjeu

As a result, data is fragmented, exchange of information is limited and effort is duplicated. The difficulty in exchanging data on food adequately and with minimum effort is a problem significant to the well-being of the human race and one that transcends political, ethnic and religious boundaries. We are all concerned with the relation between diet and disease. Epidemiologists and medical workers need dietary and nutritional data and data on food composition and biological and environmental contamination of food. These complex problems affect us all. The clues to the causes and prevention of many diseases can be revealed by comparison of dietary data between different countries.

1.2.2. Advantages

pb compatibilité
méthodes de codif

In the past fifteen years or so, the author has been concerned in the design and implementation of many food databanks for the UK Government. These included food composition tables (nutrients); packaging materials, many types of environmental contaminants such as pesticides, heavy metals and mycotoxins; additives such as flavours and colors; radio-chemical contaminants; micro-biological contaminants and so on. Some of these databanks comprised data on the composition of foods and other data collected from surveys of the diet of selected populations. In most cases, each data set employed a different method of describing (or 'coding') foods for retrieval. Each system used a description language or code of minimal simplicity designed to satisfy the immediate requirments of the scientific work and of the project initiator.

pb com

A similar situation exists in many organisations throughout the world. It means that not only cannot data be exchanged between countries, or organisations in the same country, but often that it cannot easily be exchanged between workers in the same institution. Moreover, it is difficult to link data on the same or similar foods with data in other databanks.

intérêt d'un nouvelle base
↳ pas H refaire
↳ partir de ch. communes

A common food description language could be applied to data of these types with the advantages that no new descriptive system need be developed for new databanks (thereby saving experts' time); common algorithms or subroutines could be used in computer software; databanks could be more easily merged; better use could be made of the data; and the data made use of by workers in many different disciplines and organisations.

1.2.3. Data - Organisation - Storage - Retrieval - Exchange

It is convenient to discuss these items by example - that of food composition or nutritional data.

pb ≠ pays

Well over one hundred and fifty food composition tables are listed in the INFOODS directory of food composition tables. These table present various difficulties of interpretation and use but only one concerns us here - the nature of the food to which the analyses refer. To the authors knowledge, without exception, the foods are described simply, by common names, supplemented by taxonomic Latin names where these are relevant. Occasionally descriptions include words such as 'canned' or 'frozen', 'boiled' or 'grilled'. Sometimes the recipes and cooking method for dishes are given. Foods are usually grouped according to conventional local but often irrelevant nutritional and dietary practice e.g. 'Fruits and Vegetables', 'Dairy Products' and 'Meat and Fish'. No two countries appear to agree on these classes or what should go in them, especially when mixtures are concerned. Even simple foods are difficult to classify - is a 'croissant' a cake or a bread? The information about the food is minimal although at the time of analysis it would have been comparatively easy to acquire and record.

The range of information that would enhance the value of any particular databank differs from one to another; however, briefly, food composition depends on the nature of the original commodity, how it has been grown, what has been done to it and what has been added to it. An internationally recognised method of describing these factors will have encourage recording of this important additional detail.

1.2.4. Deficiencies of Natural Language

Foods names are often inadequate and even misleading to those who are not closely acquainted with the local language and culture. Many synonyms, homonyms and homographs exist even between closely related languages such as the English spoken in the UK and the USA. The 'muffin' in the USA is close but not identical to the English 'Fairy Cake'. The USA 'English Muffin' is the nearest equivalent, but is not identical. Cuts of meat may have similar names but be different whilst the same cuts have different names. Names of what were once specific foods such as 'cheddar cheese' have become trivialised and are now used for a variety of products. Foods that are ethnic or national in origin often differ in various countries because of the necessity to comply with local regulations and consumer tastes.

Other peculiarites exist. Would all Indian or Chinese medical researchers know that 'head cheese', 'fromage du tete' and 'kase leber' are not cheeses? Indeed would a non-German European know pb/ [illegible] that kaser leber is not made from liver? Unfortunately many workers who use food data, but do not understand food, are unaware of such traps and consequently 'false drops' occur on retrieval from databanks by the use of natural language.

Even the use of recognised taxonomic nomenclature for plants and animals can be misleading, since the same or similar trivial names can be given to organisms in different species or even genus and order. Dr Wanda Polacchi as demonstrated that fish nomenclature can be particularly difficult.

Culinary and technological and terms can also be misunderstood. What is the difference between he USA 'broiled' and the UK 'grilled'? - if indeed there is a difference. How can we describe the nature of the oil used for frying within the description of the food that has been fried? How can we state whether that fried food was fresh, frozen or canned etc before cooking? In general, how can the details of growth, storage, processing, secondary storage and preparation for the table be precisely imparted so that retrieval is complete and precise?

The well-know relationship between relevance and retrieval when recalling data by means of natural language is generally acceptable where bibliographic databases are concerned. In the case of numeric databanks, it is not.

1.3. The problem more closely defined.

1.3.1. Nature of Food

Figure 1 shows the progress of food through the stages from growth to consumption together with some of the factors that can influence the nature and composition of the final product. The bottom portion of the diagram shows recursion. Foods may be stored, process, packed, stored again, processed again and repacked. At certain stages, other foods, which have traversed a similar path may be added before the final product is consumed.

Conceptually it should be possible to describe the path between sowing of the seed or animal conception by a series of alpha-numeric terms and hypercriticals (non alpha-numeric symbols). Whether or not this is practically possible or even desirable should be decided by the Task Group early in its discussions.

1.4. Possible Solutions

1.4.1. Simple Systems - Food Groups and Food Codes

pb
pas possibilité d'élargissement

This type of system has been mentioned previously. An attempt to unviersalise it has been made by Dr Lenore Arab (EUROCODE). However EUROCODE is a loosely structured listing of food names which provides no means of capturing and encoding other information.

1.4.2. Thesauri

pb Il doit suivre le standard

[ISO 2788 provides standards for the construction of mono lingual thesauri and ISO 5964 for multi-lingual thesauri. Although it is doubtful if many thesauri conform to these standards, it is suggested that every attempt be made, for any CODATA thesaurus to follow this standard].

ok : tient compte du relationnel entre [illegible]

A principal advantage of the use of a thesaurus is the ability to employ structural relationships between terms. These may be relationships of equivalence, hierarchy and association. At a higher level a thesaurus will employ relationships between sets of terms.

Various food thesauri have been constructed in the past and applied to bibliographic information. These include the CAB thesaurus, used for instance by Nutritional Abstracts, the IFIS thesaurus used for Food Science and Technology Abstracts, and the European Community Trilingual Food Thesaurus.

None of these is designed for numeric databanks and for many reasons, which will not be dealt with here, they are unsuitable for this purpose. However the terms incorporated would be an invaluble source for the design of a new system.

1.4.3. Facetted Classifications

A facetted classification divides the field into facets or sets of terms by unique and exclusive features or characteristics. It has been possible for informationm scientists to develop a theory which defines universally applicable basic categories as follows.

- entities
    - abstract
    - concrete
        - natural
        - artificial
- functions
- performers of action
    - receivers of Action
    - end products
- characteristics
- properties
    - materials & constituents
- actions
    - processes - internal
    - operations - external
- space
- time

A facetted thesaurus employs the technique of facet analysis within its construction. It is easy to see how this general concept can be adapted to describe the features of foods.

1.4.4. Entity Relationships

An entity relationship system associates things, actions, characteristics - the entities - with a defined number of relations. For instance relationships might be 'is made or comes from', and 'is'. We might then say {flour (is made from) wheat}, {bread (is made from) flour}, {wheat (is) Triticum aestivum}, {Triticum (is) Graminae}, (Graminae (is) Angiosperm} etc. In a pesticide database a relationship such as 'has undergone treatment (by)' could be used. It is evident that such a structure could be used to describe foods to a high degree of completion. A 'contains' or 'is made from' relationship solves the problem of foods containing other foods ie. {Cherry cake (contains) cake}, {Cake (is made from) flour}, etc. {Cake {is made from} hen egg}, {hens egg (comes from) hen} etc. {Cherry cake (is made from) sugared cherry}, {Sugared cherry (is made from) cherry}, {sugared cherry (is made from) sugar}, {sugar (is made from sugar cane}, {sugar cane (is) Saccharum offinarum} etc. (These relationships are merely for the purpose of demonstration).

voca
relational
model,

It is also easy to see how quantification can be incorporated eg {Chateau Ciqual 1949 (contains) ethyl alcohol, 12.2} or {standard milk (contains) total fat, 3.3} (see below).

1.5. Examples

1.5.1. FLAIR

In the early 1970s, the Ministry of Agriculture, Fisheries and Food, UK., contructed a system called FLAIR (Food Literature and Information Retrieval) based upon a broad facetted classification. Although not initially intended for numeric databanks it worked well enough to demonstrate the feasibility of that approach. It is a useful source of UK food terms and relationships

1.5.2. The FFV

Initially called the Featured Food Vocabulary but more accurately the Facetted Food Vocabulary, this system has been developed over eight or more years by the US Food and Drugs Administration (FDA) with the assistance of Prof. Dagobert Soergel (University of Maryland), again principally for indexing bibliographic material. It is however used for numeric databanks for pesticides and other contaminants as well as nutrients. Over two million records have been indexed using the system.

NB

As constructed it is a thesaural system using facetted classification but it does not conform to the ISO standard. The terms employed are very closely associated with US legislation, food supply and culture and the scope notes which qualify the thesaurus make frequent reference to the US Code of Federal Regulations.

nya ds
mots

Whilst it could be a relatively simple to task modify what are semantic deficiencies fromn the international standpoint, structurally the system leaves much to be desired. Taxonomic classification is only randomly applied, so that it is impossible to search for example, for all crucifers. There is no universal means of dealing foods which are defined by the amount of a specific constituent (eg wine or milk). It is impossible to deal with mixed foods other than by specifically including such mixtures in the thesaurus (eg carrots and peas is a term). With some exceptions, only one term per facet is allowable (although this has some advantages) and it is impossible accurately to code parts within parts of organisms. (eg since 'exudates are treated as 'parts' and only one part is allowed, 'cod liver oil' is coded as 'cod oil').

The overall rule is that the food is coded according to the major ingredient. This can lead to strange false drops and omissions in retrieval. For instance a brand of coffee drink in the USA contains sugar as the major ingredient and another contains coconut oil as the major ingredient. They are not retrievable under the code for coffee, but the latter is often retrieved when searching sugar and oils respectively. Both are retrieved as 'steeped beverages'. Similar problems occur with teas.

A major and fundamental fault with the FFV is the confusion between basic foods and their products, which are generally contained in the same facet. For example, searching by the term 'fat or oil' retrieves not only fats and oils but all products containing those fats or oils, such as soups, mayonnaises and salad creams, cream analogues and so on. Highly contrived boolean combinations with other terms using multiple 'and nots' is necessary to limit the retrieved list to the desired set. To achieve this end continuous perusal of the retrieved list must be made. In a large database this would be impractible and is a basically an undesirable practice.

It should be observed that the FDA use the thesaurus in a pre-coordinated system, indexed so as to provide the correct answers. The deficiences of the FFV are therefore less important to that organisation and indeed are in practice invisible. The construction of such inmdexing is very labour intensive. The approach is unsuitable to scientific work since it depends on pre-conceived notions of what searchers might wish to retrieve. In a post-coordinated interactive system, the FFV is extremely difficult to use, since much of the classification is apparenty arbitrary and frequent reference to the extensive documentation is required.

Nevertheless, the FFV remains a good starting point for development of a truly international and flexible facetted thesaurus, even if only the terms are used, and the structure replaced. More details are given in a separate paper. (To follow).

1.5.3. The IFDL

This acronym has been used both for 'International Food Description Language' and 'Interlinked Food Description Language'. Following meetings with foreign food and information scientists and NCI specialists, the FDA and in particular their consultant, Prof. Soergel realised the deficiencies of the FFV. As a result, the IFDL, based upon the entity-relationship model was proposed. A prototype thesaurus is currently being developed by Prof Soergal and the author, built around the concept of a PROLOG database. It is obvious to those even slightly familiar to the computer language PROLOG that the entity-relationship model briefly described above is kindred to PROLOG in format. In a databank using the system and written in PROLOG, program and data become indistinguishable. In PROLOG {is made from (bread, flour, wholewheat 95)} and {contains (milk, fat, 3.3} or {contains (potato, lysine, 340} are acceptable and embody both qualitative and quantitative statments. Units would defined within the parameters of the databank ie % for fat, mg per g for amino acids etc.

A separate paper will provide more detail of the IFDL.

1.5.4 The CSSR System.

The Food Research Institute, Bratislava, Czechoslovakia started in 1975 to design a system for use in machine readable databanks. The system extends beyond food descriptuion.

It comprises a twenty digit code used as follows.

Digit	Data
1	subject eg dishes, beverages, diets
2-3	food commodity eg eggs, meat
3	subcategories of food commodities
4-8	taxonomic information
9	state of maturity
10-13	anatomy, morphology etc
14-15	technology, processing
16-20	reserved for further use

The system is implemented by PL/1 application programs and IDMS. About 500,000 records covering scientific, technical and economic data concerning food and agriculture compiled by collaborative effort throughout the country are stored.

This is essentially a facetted system. Its effectiveness will depend upon the sub-structure ie the way terms are defined and related and can be used. The format has certain similarities to the library Dewey Classification and such methods may lead to difficulties when numbers are exhausted and new terms cannot find a place. (A UK survey of synthetic colors in foods used a similar method and experienced this difficulty). The insertion of extra digits to accomodatae the new terms can require extensive reprogramming. The system is, however, clearly sophisticated and an example of a highly developed locally applicable product. We are indebted to Dipl. Ing. Alexander Szokolay, DrSc., Director, Food Research Institute, Bratislava, Czechoslovakia for this information.

1.5.4 The INFOODS/Truswell System

This system must be mentioned since it has been widely advertised, especially to nutritionists but is not really a food language or even an information system. It is instead a listing of features or entities which might influence the composition of food, and which collectors of data should be encouraged to record. The 'thesaurus' is not a thesaurus in the technical sense, but a short, unstructured 'go-list' of prescribed words. Such endeavours have been made before without success.

1.5.5 Other Systems

In 1979 Drs. Loren Harris (Utah State University) and Ritva Butrum (then US Dept. of Agriculture now NCI) prepared a draft outline of a food databank system for the FAO Food Contamination Monitoring Program. The system was designed to suit the technology popular at that time ie 80 column card format. It listed data entities grouped together under general headings such as Country, Climate Zone, Organic Manure per unit areaclass of pesticide, parts of plants and animals and son. However there was no attempt to further structure the terms or to relate them. It contains a comprehensive lists of useful descriptors. Another system was proposed in 1981 by Drs Butrum, Sorensen and Selzer (USDA) based upon food groups, but largely unstructured. The original paper contains useful information on the availability of various aspects of information on food composition.

In the 1960s, R Jowitt (National Institute of Food Technology, Weybridge) proposed a 'dendritic' (tree-structured) system but and applied it locally.

1.5.6 The INFIC system

This system owes its origin to Dr. Harald Haendler (University of Hohenheim, Stuttgart, FDR) and to co-workers in the FDR and Utah, USA. It applies to animal feed and not human foods but is an example of a successful implementation of a facetted thesaurus ina related area. The system uses the following facets,

original material	eg plant
part	aerial part
process	fresh
stage of maturity	early vegative
cutting of crop	cut 1
grade	

A data-bank using the system is widely used throughout the world and is a reference point for all concerned in the composition of animal feeds.

1.6. Other Considerations

1.6.1. Hardware and Software

It has been assumed that any international food language will be used on machine readable data systems. In this context, it is preferable that organisations and individuals should be able to employ the system without difficulty or undue expense. Since considerable resources are necessary to develop applications software, it is preferable that the language should be useable within the framework of existing and universally available commercial or public domain software packages.

facilité d'utilisation

This restriction limits the attraction of the entity/relationship model for which only one commercially available system (in the English language) has been found. On the other hand many comparatively cheap software packages are available for dealing with thesauri.

Some of the problems noted above, for instance those concerning amounts and mixtures, might be solved within software application instead of the food language itself and experiments are being designed at NCI to investigate this possibility using commercial software.

For the language to gain widespread use and acceptibility, the software available should be useable with a variety of hardware. A number of the thesaural software packages that also deal with numbers are available on small machines have mathematical funtionality to some degree and also possess an internal macro- or programming language. There is a distinct trend towards the provision of software for mini- and mainframe computers that deals with numbers and text ie a merging of database management systems and information retrieval systems. Some of these software packages provide for multi-lingual thesauri and synonyms. Some also offer user-designed links to external application programs and user written subroutines. Links to other software systems can be envisaged.

ordinateurs

Additionally many are already available in read-only mode for use with CD-ROM. This affords the possibility of distibution of both software and large volumes of data to research workers for use with micro-computers.

1.7. Summary

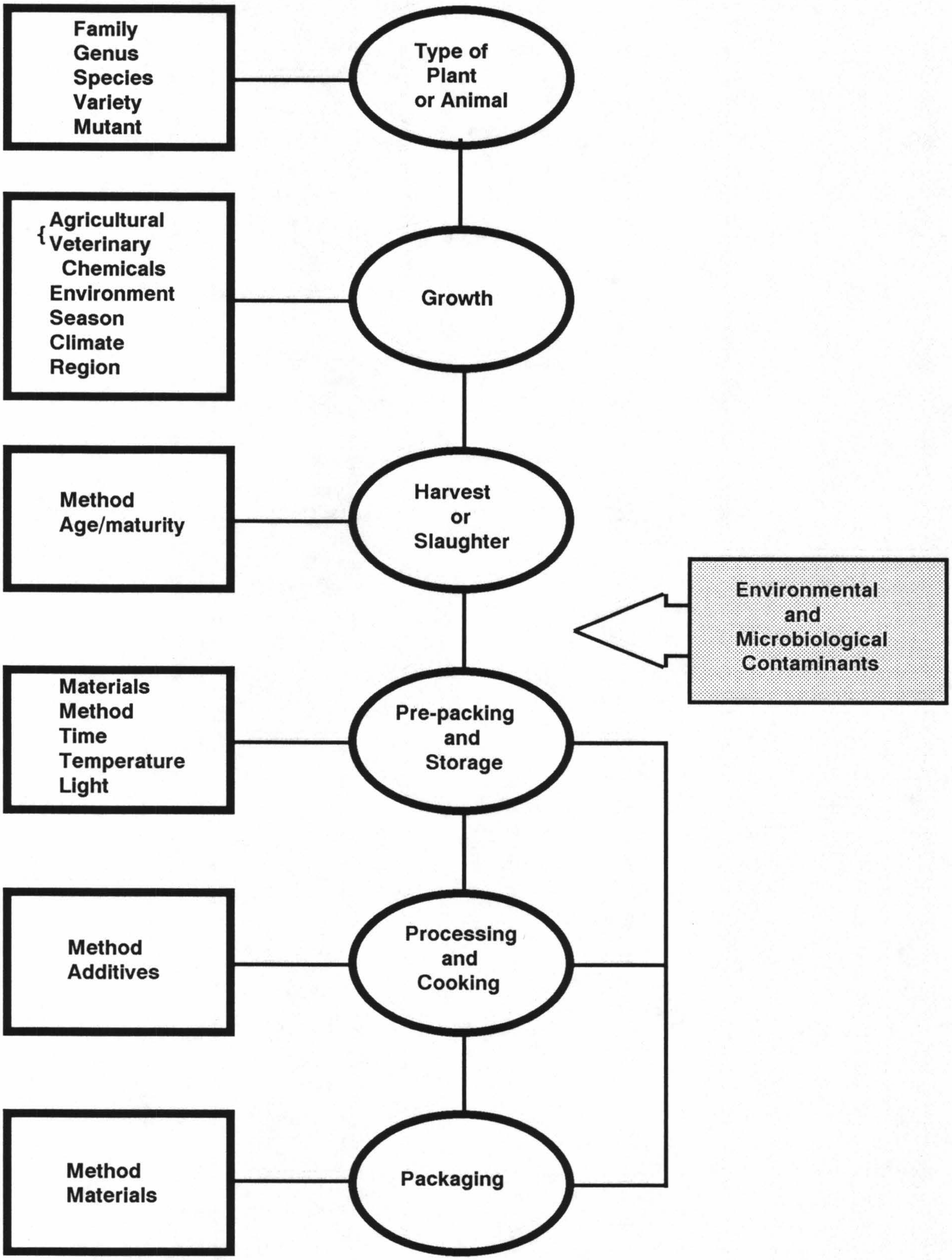
enjeu

1.7.1 A universally applicable method of describing or coding foods for use in numerical databanks would facilitate the improved use and interchange of data.

à faire

1.7.2 A methodology for such an international food language must be made. A conventional facetted thesaurus, free from national, cultural or ethnic bias would be difficult to construct but might present less overall problems than alternatives. The entity-relationship model needs further investigation. However any other system should also be considered.

Factors Affecting Food Composition


Family
Genus
Species
Variety
Mutant
Type of Plant or Animal
{Agricultural
Veterinary
Chemicals
Environment
Season
Climate
Region
Growth
Method
Age/maturity
Harvest or Slaughter
Environmental and Microbiological Contaminants
Materials
Method
Time
Temperature
Light
Pre-packing and Storage
Method
Additives
Processing and Cooking
Method
Materials
Packaging