Harald Haendler
Leinfelden-Echterdingen

Synthetic Description Systems for Accurate Data Identification and Selection. Principles and Methods of Nutritional Data Banks

Harald Haendler: **Synthetic description systems for accurate data identification and selection. Principles and methods of nutritional data banks.** In: Int. Classif. 15 (1988) No. 2, p. 64–68, 14 refs.

The peculiarities of numerical data as stored in and retrieved from nutritional data banks require highly efficient systems for data identification. The principles which have to be considered for establishing such systems are demonstrated by the methods applied in the International Network of Feed Information Centres. Most important is the principle of (systematic) synthesis, which has been used at two levels: (1) to synthesize the entire informeme/data unit by its constituents which are the object, the property, and the conditions; and (2) by using a faceted description system (thesaurus) for the identification of feeds as the object. The last one is based on semantic factoring of feed denominations, whereby the analysis/synthesis follows strictly the structural relationship of the semantic components, which are identical with the essential characteristics of the object, and which belong to the six categories/facets: Origin, Parts, Process, Growth stage, Cut, and Grade. In addition to the description with descriptors of these facets, special identification systems are used for the properties (substances/values), and for the "individual" characteristics of the sample to enable a very precise selection of data units for different purposes. (Author)

1. Introduction

Information to be stored in numerical data banks may be of the kind as shown in the following three examples:

(1) The altitude of Mount Everest is 29028 feet above sea level.
(2) The melting point of lead is 327 degrees Celsius.
(3) The protein contents of wheat flour is 12.85%.

The propositions or informemes (as they are better called when used as units in the information process[1]) represented by these sentences have something in common. In each of these cases we can identify

- a special object (Mount Everest, lead, wheat flour),
- a property (altitude, melting point, protein contents),
- a numerical value for the property (29028 m, 327 degrees 12.85%).

The language representations of such informemes (usually expressed in a standardized form) are the data units which can be processed in the data bank and retrieved from it separately.

Data units of this kind play an increasing role in research and decision making in virtually each field of knowledge and in practical life[2]. A prime role in our world is played by nutrition. Overcoming the world-wide problem of hunger and malnutrition, of preserving life and health as far as they depend on a sufficient supply of nutrients requires an immense amount of reliable nutritional data, especially on the composition of animal feeds and human foods[3].

The example (3) given above shows the special type of data required for this purpose. A comparison between this example and the other two shows that the object of (3) – wheat flour – is neither an individual object (like Mount Everest) nor a well-defined chemical element (like lead), but a heterogeneous material which may differ in its composition due to biological, economical, or technical influences.

The original feed or food composition data (as produced in chemical laboratories) have to be treated in different ways to generate reliable information. The data often has to be selected and combined for special calculation (for instance, for the calculation of averages or deduced values). If irrelevant data are included in such a set of selected data, the deduced data will be wrong. Therefore, to use this kind of data successively, the object in question has to be identified and described in a very accurate manner in order to permit the necessary differentiation of the material.

These conditions have challenged the compilers of nutritional data banks and networks (or those responsible for them) to create highly effective systems for data identification. It is remarkable in this connection that completely independent working groups have arrived at the same principles as being essential for achieving the wanted degree of accuracy. It may be remarkable, too, that often different terms are used by some of these groups to name these principles, which shows that they perceived the theoretical basis without having had any prior knowledge of it from the literature.

It is the intention of this paper to treat these principles as found by the system developers in the light of accumulated experience and reflection, and to explain the methods to which these reflections have led. This demonstration will be based on the system developed for the International Network of Feed Information Centres (INFIC)[4]. The principles and methods of this system, now having been in use for about fifteen years, are very similar to those of some systems established later (or still being developed) for describing data on the composition of human foods.

2. Synthesis and Structure

One of the main (if not *the* main) steps in the development of classification and indexing is the consideration of the principle of synthesis or composition, an insight for which we are indebted to J.O. Kaiser and S.R. Ranganathan[5]. This principle opened the way from convention to creation, from stagnation to progress, from limitation to infinity. Incidentally, the composition of things that have exist for a long time, but (perhaps) have never been put together before, is an act of creation in

a very general sense: in biology (esp. genetics), in arts, in sciences, in architecture, in drawing conclusions, and in using language[6].

It is obvious that this principle of synthesis must be considered as an essential prerequisite for a highly effective description system for data on feeds and foods. It is equally obvious that the synthesis must be carried out in a systematic way, since each synthesis has to follow rules. Though the elements to be combined may vary, the general plan according to which an element of a special category can be placed within the whole is fixed. This applies to all kinds of compositions. How the constituents of a composition are related to one another constitutes the structure of the whole.

The structure of a unit of numerical data is shown in examples (1), (2), (3): The constituents are at least a special object and a property, including a numerical value for the latter. Sometimes there are still other constituents in data units. As mentioned before, there is a difference between examples (1) and (2) on the one hand and (3) on the other. The values given in (1) and (2) are independent of any condition; the altitude of Mount Everest and the melting point of lead should always be the same. But this is not true for (3) or other data units of this type. Another very simple example may show this:

(4)The boiling point of water is 100 degrees Celsius.

This statement (the given value) is only then correct if the measurement is carried out near sea level, but would be wrong for instance at the summit of Mount Everest. That means: the validity of a value may depend on specific conditions. This is often the case with analytical values of feeds and foods. It has been found necessary, therefore, to consider the specific condition as well as a constituent of the data unit. Since each constituent has to be identified in an adequate manner, the feed data bank system comprises three different identification systems, each for one of the three constituents of a data unit:

- for the object, i.e. the feed,
- for the property, i.e. the nutrients (substances) and values,
- for the conditions, i.e. the pecularities of the feed sample.

3. Faceted Object Description

The identification of the object requires great care. A source of additional difficulties for the international feed data bank system was the fact that many of the usual feed denominations are synonyms or polysemes, with the meaning of a term (even within the same language) often varying from region to region. Another problem is the immense and still increasing number of feeds for which data have to be stored.

The solution to both these problems has been found in using the synthetic principle for the description of feeds as well. To make an effective use of this principle in this case, each feed denomination has first to be analysed semantically to determine its semantic components whereupon a language representation for this component has to be chosen or to be coined as its descriptor. This method is known as semantic factoring which may be the reason why a similar system for describing human foods has been called "Factored Food Vocabulary"[7].

It should be mentioned in this context that the useful semantic components of feed denominations are precisely the characteristics of a feed which determine its specific kind and value and which, therefore, are also useful for describing typical sets of selection.

Another important aspect is the internal structure, i.e. the kind of relationship existing between the semantic components. This relationship depends on the kind of the characteristic in question or, what is the same, the category it belongs to. The semantic components which together determine a specific feed belong to different categories or, as we could call them in this context, facets. Semantic analysis has to consider these facets systematically. Later on the synthesis of semantic component descriptors for defining a feed of which data should be selected must be carried out in an analogous way. This is what Kaiser called systematic indexing as far back as 1911.

The categories to which already Aristotle assigned all concepts (or terms) used in his time are divided into a "basic category" of substances (or entities, objects, concretes etc.) on the one hand, and categories of attributes, properties, processes etc. on the other hand. The necessity to differentiate between "concretes" and "processes" in systems for systematic indexing was already pointed out by Kaiser. A description system for objects like feeds must of course first consider the fundamental concrete material, which later can be specified by adding further characteristics in the form of descriptors taken from the other facets.

The facets of the system for feed description follow these principles. The "concretes" in the case of a feed is the original material the feed consists of or is produced from. Since such materials as plants, animals, minerals or artifacts are not always used completely, the special part used as feed has to be identified. Hence, it is necessary to consider the two facets (Fc.):

Fc.1: ORIGIN and
Fc.2: PART.

For instance the object in example (3) has two different components belonging to these two facets: "wheat" is the full plant (belonging to Fc.1), "fluor" is a part of this plant, more precisely a part of a broader part "grain", both these parts belong to Fc.2. Other examples for part descriptors are:

(5) aerial part/leaves/germs/bran

The materials usually undergo special processes or treatments, which form a further kind of characteristics to be described by a descriptor of the facet:

Fc.3: PROCESS.

Examples of descriptors of this facet are:

(6) ground/germinated/cracked/toasted/dehydrated.

These three categories of characteristics must be considered in each case. Other categories may be applicable

in special cases. For those cases the system provides the following three facets:

Fc.4: GROWTH STAGE
for identifying the stage of maturity of plants or animals,

Fc.5: CUT
for green forages like grass, clover etc. which may be cut and used severel times during the year,

Fc.6: GRADE
for special factors not identifiable by descriptors of Fc.1 through 5 which may have influenced the quality or composition of the feed.

4. Semantics and the Thesaurus

The application of the descriptor vocabulary, systematically structured in facets, allows the description of a feed in a correspondingly systematical way by enumerating the correct descriptors for its characteristics. This description is unambiguous insofar as the descriptors used, are in themselves unambiguous. In case the terms are polysemes, monosemy must be achieved by artificial devices.

Using semantic factors as descriptors rather than complete denominations is already an effective method of monosemysing, because the enumeration of the characteristics explains the content of the denomination. This is really a semantic analysis. But nevertheless some terms, apparently able to represent a characteristic may be polysemic themselves, so that their use may lead to ambiguous descriptions. To avoid this, all "candidate descriptors" had to be tested thoroughly to prevent polysemy within the vocabulary. Furthermore it was necessary to display the vocabulary in such a manner that the correct use of all descriptors is facilitated. That means the vocabulary had to be developed into a thesaurus which is called the "INFIC FeedThesaurus"[8].

The assignment of a descriptor to one of the six facets gives a first information about its meaning. A descriptor's belonging to a specific facet determines whether it stands for the original material, the part, or the process; this may not always be clear from a descriptor in isolation. The composition of descriptors from different facets in a prescribed range is an effective syntactical device, like the grammar of natural languages, which devides the constituents of a sentence into subject, predicate, object, etc.

The choice and coining of descriptors raised some other problems, especially within Fc.1. The large number of plants and also animals which can be used as feeds, and the multiplicity of denominations used in different regions and language communities for the same thing (synonyms) required extreme thoroughness. Therefore it was decided to use the scientific (Latin) denomination with genus and species and, if applicable, with variety. Additionally a preferred term of the common (trivial) denominations is used.

Descriptors of this facet are for instance:

(7) Triticum aestivum, wheat;
(8) Citrillus vulgaris, watermelon.

Another problem concerning "descriptor making", not only in the case of Fc.1 but of all facets, was the multiplicity of natural languages. The difficulty lies in the fact that corresponding terms of different languages are not always semantically equivalent, their meanings differ from language to language as they are polysemic in different dimensions[9]. Because the INFIC FeedThesaurus is a tool for an international network, it was necessary to cope with this problem of insufficient semantic equivalence between natural languages. A first step was the reduction of the languages of the vocabulary, i.e. the languages in which descriptors are provided, to English, German, and French (which are the "system languages"). This does not mean that other languages are neglected, they have been considered for delivering lead-in terms.

Within the „system languages", all efforts have been made to achieve strictly semantic equivalence between corresponding descriptors by choosing useful terms with clear meaning in all languages, or by using qualifiers to restrict, if necessary, the meaning of terms in one language, or by explaining to what extent the descriptor should be used, i.e. by giving a scope note.

Like other thesauri the INFIC Feed Thesaurus displays relationships. The generic relationship between descriptors within the same facet is a very effective means to make the conceptual structure transparent, in which connection it is helpful to ensure the correct use of a descriptor. Apart from the generic relationship, the partitive relationship sometimes, plays a role also.

The associate relationship is likewise included in the thesaurus. But contrary to general usage the INFIC Feed Thesaurus strictly distinguishes between, on the one hand, "real" related terms in the sense that both terms are descriptors within the same facet (i.e. a paradigmatic relationship), and, on the other hand, those terms which are often used in combination within the same description. In the latter case they tend to appear together in contexts, although belonging to different facets (i.e. a syntagmatic relationship). Only in the first case are the terms called related terms (RT), while in the second case it is better to call them "affinitive terms" (AT). It was felt that the distinction between these different relationships is not only a helpful device for a precise description, but also a logical prerequisite for a faceted system.

Different from the relationships among descriptors are those between descriptors and non-descriptors, i.e. among synonyms. The "used for" – reference is an additional information given to the descriptors referring to those synonyms which are not chosen as descriptors. The most important reference is the "use" – reference which leads from the non-descriptor to the descriptor. A peculiarity of the INFIC Feed Thesaurus is that for Fc.1 vernacular names of plants as used for instance in tropical countries are considered as non-descriptors from which reference ist made to their respective descriptors. This allows the identification of a feed in case only local denominations are available. These terms not belonging to one of the "system languages" are listed in an auxiliary part of Fc.1 serving as a reference tool or lead-in vocubulary.

It may be evident from these various additional devices that the INFIC Feed Thesaurus is a very sophisticated instrument for achieving an unambiguous and very correct description of a feed independent of local usage of terms and of other linguistic difficulties. The thesaurus is also a flexible system for describing any new feeds with the existing descriptors by setting up new compositions of these descriptors.

Up to now about 25000 different feeds have been described by this system. All these descriptions are recorded in the "International Feed Description File"[10] so that no new description is necessary if the same feed has been described before. This list is serially numbered, and the "International Feed Number" (IFN) can be used for identifying a feed. The IFN can also be used for a language-independent identification within the network, especially for computer operation.

Examples of items of the "International Feed Description File" (IFN and descriptions) are as follows:

(9) 5–29–796 Triticum aestivum, wheat, malt sprouts, dehydrated.
(10) 5–27–719 Citrillus vulgaris, watermelon, seeds oil residue, mechanical extracted ground.

5. The Coding System for Substances and Values

It was mentioned before that the identified feed is only one constituent of the whole informeme represented by a data unit. At least one property of this feed has to be included to obtain a proposition which can be used for information. Therefore, besides the identification of the feed, it is necessary to identify each property which has resulted from a laboratory analysis of this feed. Although this identification is not as difficult as that of the feeds, it must be done very carefully and in such a way that precise retrieval under the aspect of the property is possible.

Since the establishment of the feed data bank system the number of properties or kinds of values has increased steadily. At the beginning of the feed data bank activities only "classic" nutrients and their digestibility coefficients were stored. Later on, minerals and trace minerals, amino acides, vitamins, fatty acids, and the multitude of diverse toxic substances were included in the kinds of properties of which values have to be stored in the data bank. For this purpose a special coding system was developed. The codes consist of a three digit number so that up 999 different codes for the kinds of values can be used. The system also includes codes for fixing the weight unit or how the concentration of a substance is recorded. Furthermore the analytical method used by the laboratory and others can be identified by assigning the respective codes. Some of these codes can also be used as factors for converting the data into a standard form to make them comparable and usable for specific purposes.

Codes (so called "attribute numbers") for substances with the "unit" of which the data have to be recorded are for instance:

(11) 109 protein %
(12) 652 riboflavin mg/kg.

6. The System for Sample Description

As mentioned before, the validity of a value may depend on specific conditions which the feed sample has undergone. With an increasing number of substances (others than the classic nutrients) that had to be considered important for nutritional information the question arose whether the value could be assessed as "typical" or "non-typical". For correct use of the data it became increasingly important to know what environmental factors might have influenced the concentration of a substance in a feed sample. For this reason a special system for describing pecularities of the feed sample has been established, allowing the identification of kinds of environmental factors as well as their degree.

With the inclusion of such a recording system for "secondary information" a distinction between, on the one hand, a feed as the class of all comparable feeds and, on the other hand, the individual sample was set up. The difference between "class" and "individual" (as one element of this class) is clearly demonstrated by the two description systems: All characteristics which can be described with descriptors of the six facets of the thesaurus are essential characteristics of this class of feeds. But everything for which an additional description is required pertains only to the individual sample, i.e. these are individual characteristics not typical of the whole class. This distinction is important for retrieval purposes. In general, averages may be calculated from all elements of one class. But where individual characteristics of the sample are concerned it is not be permissible to compile all elements of the class for averaging.

The special system for sample description is very voluminous. It comprises codes for numerous environmental factors like traffic, industry emission, fertilizing, pesticide application, soil quality, storage conditions, weather conditions, and others.

For instance the following codes are used to identify pollutions sources:

(13) 022 = potassium industry and saline.
(14) 070 = nuclear energy production plants.

The distance between the source of emission and the receptor (the place where the sample is taken) is being recorded by additional codes like

(15) 01 = 0 – 10 m
. . .
10 = 501 – 700 m.

In addition to this description of the specific factors of the sample, all formal information about the region or country, the laboratory, or bibliographical details in the case of publication of the data, is systematically recorded. This makes it easier to evaluate the data and, if necessary, to contact the data producer for ensuring that only reliable data are stored so as to provide users with reliable nutritional information.

In *conclusion* it can be maintained that the multiple possibilities to synthesize different kinds of identification elements (descriptors as well as codes) allow, together with an adaquate software of course, the selection and compilation of accurately reliable nutritional data in many diverse formats directly from the data bank.

Notes:

1 Concept and term "informeme" was introduced by A Diemer, see (1).
2 Cf. Haendler in (10).
3 Cf. Haendler in (6).
4 Cf. Hacndler in (3) and (4).
5 We are indebted to E. Svenonius – cf. (14) – for recalling Kaiser's ideas of systematic indexing published as far back as 1911 – cf. (13) – i.e. before Ranganathan's Colon Classification.
6 The creative aspect of language usage and its significance for documentary languages has been treated with reference to W. v. Humbold and N. Chomsky in (2).
7 The "Factored Food Vocabulary" (FFV) has been established by the US Food and Drug Administration. The Vocabulary will be developed into a food identification systems for broader use by the National Cancer Institute, Bethesda MD/ USA with assistance of an internatinal working group.
8 The "INFIC Feed Thesaurus", consisting of several voluminous parts, is not printed as a whole, but is in use in the form of computer files and print-outs; the thesaurus or some aspects of it has been repeatedly described, cf. (5), (7), (8), (9), and (12).
9 Cf. Haendler in (11).
10 Cf. Harris et al. in (12).

References

(1) Diemer, A.: Informationswissenschaft. Zur Begründung einer eigenständigen Wissenschaft zur Grundlegung eines autonomen Bereichs "Informationswissenschaften". Nachr. Dok. 22 (1971) No. 3, p. 105–113.
(2) Haendler, H.: Selektionsgerechte Indikation von Sachgebieten und Sachverhalten (Selection-oriented indication of subject fields and propositions). In: Int. Classif. 2 (1975) No. 1, p. 25–31.
(3) Haendler, H.: International cooperation in data documentation of feed analyses. In: Proc. Fifth World Congress of the Int. Assoc. of Agricultural Librarians and Documentalists, Mexico City 1975, p. 18–22.
(4) Haendler, H.: Feed data banks and the International Network of Feed Information Centres. In: Proc. Symposium organized by the Commission of the European Communities, Luxemburg 1977, p. 18–22.
(5) Haendler, H.: The INFIC Feed Thesaurus. In: Proc. INFIC Workshop "Up-Grading International Feed Data Bank System", London, UK, 1981. p. 35–42.
(6) Haendler, H.: The accessibility of nutritional data – necessity and realisation. In: Proc. Eighth Int. CODATA Conference, Jachranca, Poland 1982. Amsterdam: North-Holland p. 321–324.
(7) Haendler, H.: Konzipiert für die Belange der Datendokumentation: der Internationale Futtermittelthesaurus. In: Studien zur Klassifikation 14 (1985) p. 167–174.
(8) Haendler, H.: Methods of identifying data units for retrieval purposes, as applied in an international data bank system for feed analyses. In: Proc. International CODATA Conference, Jerusalem 1985. Amsterdam: Elsevier Science, p. 401–404.
(9) Haendler, H.; Neese, U.; Jager, F.; Harris, L.E.: INFIC Feed Thesaurus, a multilingual thesaurus for describing feeds for the data bank of the International Network of Feed Information Centres. Prepared on behalf of INFIC by Dokumentationsstelle der Universität Hohenheim, Stuttgart, Germany, F.R.
(10) Haendler, H.: Analytical data as information sources. In: Quarterly Bulletin of the IAALD 32 (1986) p. 159–162.
(11) Haendler, H.: Probleme der inhaltlichen Erschließung von Dokumenten in internationalen Verbundsystemen. In: Mitteilungen der Gesellschaft für Bibliothekswesen und Dokumentation des Landbaues 38 (1986) p. 59–84.
(12) Harris, L.E.; Jager, F.; Leche, T.F.; Mayr, H.; Neese, U.; Kearl, L.C.: International feed descriptions, international feed names, and country feed names. International Network of Feed Information Centres Publication No. 5, prepared on behalf of INFIC by the International Feedstuffs Institute, Utah State University, Logan, Utah: 1980. 11 + 769 p.
(13) Kaiser, J.: Systematic indexing. The card system series II., 1911 J. Gibson, London, UK.
(14) Svenonius, E.: Facet definition, a case study. In: Int. Classif, 5 (1978) No. 3, p. 134–142.

Address:
Dr. Harald Haendler, Viehweg 19,
D-7022 Leinfeldcn-Echterdingen.

LC Subject Authority File Available on CD-ROM

CDMARC Subjects, the complete subject authority of the Library of Congress on CD-ROM, is a reality. It is now available for use on IBM PCs or compatibles by subscription with three isues in 1988 for USD 300 (U.S.) and USD370 (international).

CDMARC Subject offers a hierachical approach to subject headings, making it simple to maneuver intelligently through layers of interrelated subject terms. It is the first of three CD-ROM products from CDS, the Cataloging Distribution Service of the Library of Congress. Work is continuing on CDMARC Names, LC's complete name authority product and CDMARC Bibliographic, the complete file of LC's MAchine-Readable Catalog (MARC) records for books, serials, music, maps and visual materials. For more information on CDMARC products, write to CDS at the address: Library of Congress, Customer Services Section, Cataloging Distribution Service, Washington, DC 20541, USA. (From CDS Connection (1988)No.2)

CDMARC Subjects was to be the featured item on this year's agenda at the Congress of the American Library Association in New Orleans, July 12, 1988. Experience with this newest product was to be discussed by organizations that have actually used it as part of a nationwide testing program.

Library of Congress Subject Headings (LCSH) 11th Edition

LCSH 11 has been published recently as a three-volume set. It contains all additions and changes to LC's subject authority file cumulated through the first week in September 1987.

The publication is now issued annually instead of every four years. It has a completely revised subject headings format, making the headings more understandable and easier to use. Much of the material formerly included in the "Library of Congress Subject Headings: A guide to Subdivision Practice" has been included in LCSH 11 in the form of scope and reference notes. The balance of the material from the guide may be found in the "Subject Cataloging Manual: Subject Headings". Quarterly and annual supplements have been discontinued, but the weekly tape service, quarterly microfiche, and paper weekly lists will be produced. The price of the new three-volume set is USD150. (CDS Connection (1988)No.2, p.2)

Classification Broadside

The twenty broad subject areas and subareas of the Library of Congress Classification System have been printed on two 11x17-inch posters and are available now as the "LC Classification Broadside". They may be placed near card catalogs or shelves of books as a helpful guide for library users. The price is USD 15. (CDS Connection (1988)No.2, p.2).