Francis J. Devadason
DRTC, Bangalore, India

Computerization of Deep Structure Based Indexes*

Devadason, F.J.: **Computerization of deep structure based indexes.**
Int. Classif. 12 (1985) No. 2, p. 87–94, 26 refs.
The Deep Structure Indexing System is based on: 1) a set of postulated Elementary Categories of the elements fit to form components of names of subjects; 2) a set of syntax rules with reference to the Categories; 3) a vocabulary control tool such as Classaurus; 4) a set of indicator digits to denote the Categories and their subdivisions; and 5) a set of codes to denote a few of the decisions of the indexer. Names of subjects formulated on the basis mentioned above are input to a set of computer programs to generate several different types of subject index entries. This paper describes briefly the Deep Structure Indexing System. (Author)

1. Subject Indexing Language

A Subject Indexing Language (SIL) is an artificial language used for formulating names of subjects. "The name of a subject is essentially a piece of non-discursive information, and it is an indicative formulation that summarises in its message 'what a particular body of information' or document is about" (1, 2, 3). The primary components of a SIL are: 1) the elements or vocabulary; 2) the categories of the elements on the basis of their semantic significance; and 3) the rules of syntax with reference to the categories, for formulating admissible names of subjects, indicating the nature of relationship between the elements (4, 5).

2. Deep Structure Indexing System

In the name of a subject three different structures namely, semantic, elementary and syntactic can be recognised (1, p 91; 2, p 12; 6). By logically abstracting the structures of SILs of Kaiser, Cutter, Dewey and Ranganathan a 'Deep Structure of SILs' (DS of SIL) has been arrived at by Bhattacharyya (2, p 12; 7). The Deep Structure Indexing System (DSIS) is based on DS of SIL. It is based on: 1) a set of postulated Elementary Categories of the elements fit to form components of names of subjects; 2) a set of syntax rules with reference to the categories; 3) a vocabulary control tool such as the Classaurus; 4) a set of indicator digits to denote the Categories and their subdivisions; and 5) a set of codes to denote a few of the decisions of the indexer, in order to generate by computer manipulation, different types of subject indexes.

* Paper presented at 9th Annual Conference, Gesellschaft für Klassifikation eV, Karlsruhe, FRG, 26–28 June 1985.

2.1 Elementary Categories (ECs)

The DS of SIL postulates that the component ideas in the names of subjects can be deemed to fall into any one of the ECs: Discipline (D), Entity (E), Property (P), and Action (A). In other words, a component idea can be a manifestation of only one of the ECs (2, 6).

Note: The term 'manifestation' has been used to denote an 'instance' or 'occurrence' of the respective ECs. The names of ECs are also used to denote their respective manifestations in this paper.

2.2 Subdivisions of Manifestation

Manifestations of each of the ECs may admit of subdivisions: Species/Type, Part and sometimes Constituent. A Species/Type does not disturb the conceptual wholeness of the manifestation to which it is a Species/Type. A Part is a non-whole of the manifestation to which it is a part. A Constituent is an ultimate part with its own individuality. For example, in the case of "House", 'Multi-Storeyed House' is a Species/Type; 'Foundation', 'Roof', 'Door', 'Window' are Parts; 'Cement', 'Mortar', 'Sand' are Constituents (8). Constituents generally occur for the EC Entity.

2.2.1 Modifier: Compound and Complex Terms

Apart from the ECs, a special component called Modifier is also recognized in the names of subjects. Modifier is an idea used or intended to be used to qualify (differentiate, speciate) the manifestation without disturbing its conceptual wholeness. For example, 'Red' in 'Red Rose'; 'Concrete' in 'Concrete Bridge'; 'Infectious' in 'Infectious Disease'. A Modifier generally creates a Species/Type of the modifyee (focus). Modifiers can be Common Modifiers like Form, Time, Environment and Place, and Special Modifiers which can be based on any of the ECs. Generally Common Modifiers can modify a combination of two or more manifestations of two or more ECs.

Depending on the structure of the 'Modified Term', Modifiers could be further grouped into two types:

1) Modifier of Kind 1, that which requires auxiliary/function words to be inserted between the modifyee term and its modifier term forming a Complex Term (Ex: 'Finishing *for* Tip Effect' which is a type of 'Finishing' of Leather); and
2) Modifier of Kind 2, that which does not require auxiliary/function words to be inserted in between, but automatically forms an acceptable Compound Term (Ex: 'Foot' forming the Compound Term 'Foot Bridge' which is a type of 'Bridge').

Note: The above grouping of modifiers depend on the natural language used for indexing. Moreover, that component in the name of a subject represented as a Complex Term is likely to be changed into a Compound Term/term by subsequent emergence of new technical terms in the subject area concerned. The auxiliary/function words in Complex Terms may be role indicating words or 'phase relation' indicating words or prepositions etc. These auxiliary/function words may have to be standardized to achieve consistency in the representation of names of subjects.

2.2.2 Composite Term

According to the postulate of ECs (See sec. 2.1), any one component in the name of a subject can belong to any one (only one) of the ECs. If a component term represents manifestations of more than one EC then it is a

Composite (Category) Term. It should be broken down (factored, decomposed) into two or more constituent terms and each one of them should be identified as belonging to one or the other of the ECs.

Perhaps Kaiser is the first person to realise this (10) and Ranganathan used the idea as a basis for the development of faceted classification schemes (11, 12, 8 p 404–5). The identification and factoring of Composite Terms is guided by the ECs of the indexing language (13). The Composite Term is considered in DSIS as a synonymous term to the combination of the factored constituent terms. For ex: Phthisis = Medicine (D) + Lung (E) + Tuberculosis (P).

Note: There are two other concepts viz., 'Base' and 'Core' which are not discussed in this paper.

2.3 Syntax of DS of SIL

The basic rule of syntax associated with the DS of SIL for formulating names of subjects is that, Discipline should be followed by Entity (both modified or unmodified) appropriately interpolated or extrapolated wherever warranted by Property and/or Action (both modified or unmodified). The other rules of syntax have been discussed by Bhattacharyya (2, 6, 7, 9). In general, the rules of syntax give rise to the following sequence of components in a name of subject:

DISCIPLINE followed by ENTITY which is followed by PROPERTY and/or ACTION. PROPERTY and/or ACTION may be further followed by PROPERTY and/or ACTION as the case may be. Each of the above manifestations may further admit of, and be followed immediately by their respective SPECIES/TYPES and/or MODIFIERS and/or PARTS and/or CONSTITUENTS. The COMMON MODIFIERS generally occur last in the sequence.

The rules of syntax give rise to a context-dependent sequence of the components in the name of a subject in conformity with Ranganathan's principles for facet sequence such as the Actand-Action-Actor-Tool principle (8, p 425–9).

2.4 Indicators of Deep Structure

Certain numeric codes have been prescribed in DSIS to indicate the manifestations of the different ECs, their subdivisions and modifiers of different kinds. These indicators are given below:

Common Modifiers	*Elementary Categories*
0 Form Modifier	9 Discipline
2 Time Modifier	8 Entity
3 Environment Modifier	.2 Property
4 Place Modifier	.1 Action

Subdivisions/Divisors
.3 Constituent
.4 Part
.5 Modifier of Kind 1 including Phase Relation Modifier
.6 Species/Type including those created by Modifier of Kind 2

In the name of a subject the indicators precede the components to which they are indicators. The indicators for Property and Action and also for the Subdivisions/Divisors are attached with the indicators for the ECs to which they are respectively Property or Action or Subdivisions/Divisors (See sec. 3.5).

3. Formulation of Name of Subject

3.1 Expressive Name of Subject

Taking the title as the starting point, names of each of the specific subjects dealt within the concerned document are expressed in natural language. Each of the component ideas corresponding to each of the ECs that are implied, are explicitly stated in each of the names of subjects to form an 'expressive title'. Let one of the expressive titles be:

"In Leather Technology, Evaluation of Two Bath Chrome Tanning of Leather using Dichromate".

3.2 Formalised Name of Subject

The expressive title is then analysed to identify the ECs, their subdivisions/divisors, to which each of the components in the expressive title belong. All Composite Terms are factored into their fundamental constituent terms and identified as belonging to one or the other of the ECs. The Composite Terms are noted separately for preparing Cross Reference (CR) entries to be included in the final index. The component terms are written down as a formalised expression following the rules of syntax, as given below:

"(Discipline) Leather Technology, (Entity) Leather, (Action on Entity) Two Bath Chrome Tanning, (Entity based Modifier of Kind 1) *(Using)* Dichromate, (Action on Action) Evaluation".

3.3 Modulated Name of Subject

Each of the component terms in the formalised name of subject is then analysed to find out its superordinate terms. This is done by finding out "of which the concerned component is a Species/Type or Part or Constituent?", in the context of the name of subject as a whole. This process is continued with each of such superordinates recognised in the process till it ends up with the concept of the EC of which it is a manifestation. For this purpose terminological sources such as thesauri, classauri (14, 15), dictionaries etc., are used. Each of the superordinates are fixed prior to the concerned term successively giving rise to a 'Modulated' name of subject as follows:

"(D) Leather Technology, (E) Leather, (A) Tanning, (Type of A) Mineral Tanning, (Type of A) Chrome Tanning, (Type of A) Two Bath Chrome Tanning (Modifier of Kind 1) (using) Dichromate, (A on A) Evaluation".

Note: The reason for making each of the superordinates to precede the respective component terms is to endow the name of subject with the capacity to produce an organising sequence effect resembling the sequence of class numbers. Moreover, it is possible to prepare the alphabetical subject index using all or the relevant superordinate terms as Lead Terms, from the name of the subject itself. Generally it is not necessary to 'modulate' Modifier of Kind 1, forming a Complex Term in the name of a subject.

3.4 Standardized Name of Subject

Each of the component terms in the name of a subject is replaced with Standard Terms and the synonymous, quasi-synonymous terms are noted separately for preparing CR entries to be included in the index later. For this purpose vocabulary control tools such as thesauri, classauri etc., are used.

3.5 Name of Subject with Indicators

Appropriate indicators for ECs, their subdivisions/divisors and Common Modifiers of different kinds are inserted in the appropriate places. The auxiliary/function words introducing Modifiers of kind 1 are also standardized, if found necessary. The resulting name of subject is as follows:

"Leather Technology 8 Leather 8.1 Tanning 8.1.6 Mineral Tanning 8.1.6 Chrome Tanning 8.1.6 Two Bath Chrome Tanning 8.1.5 (using) Dichromate 8.1.1 Evaluation".

Note: The indicator digit for Discipline is not used as it is taken as understood to be the first digit in all names of subjects. In the component '8 Leather', the indicator '8' denotes that it is a manifestation of Entity. In the component '8.1 Tanning', the indicator '8.1' denotes that it is an Action on Entity, and so on. A set of 'Modulated' names of subjects with appropriate indicators when just sorted alphanumerically can produce an 'organizing classification effect'. This has reduced considerably the *See also* CRs from narrower subjects/terms to their respective broader subjects/terms (ascending references) and from broader subject/terms to their respective narrower subjects/terms (descending references). (See also sec. 9). (See also Sec. 2.3 in reference 23).

4. Subject Index Entry

4.1 Functions of a Subject Index Entry

A subject index entry consists of index terms denoting a subject along with its address. Index terms are generally components of names of subjects. Subject index entries in general, perform three functions (16): 1) locating function – to permit the location of entries for the subject sought; 2) comprehending function – to give data for the comprehension of the entries to permit relevance judgement; and 3) organizing or relating function – to help locate entries for subjects related to the one being sought.

4.1 Structure of a Subject Index Entry

A subject index entry consists of a Lead Heading at its beginning with a Lead Term occurring first in it. The Lead Term caters to the 'locating function'. The Lead Heading may contain terms other than the Lead Term, specifying (qualifying) the Lead Term. In DSIS the Lead Headings do not contain any indicators for the ECs and their subdivisions/divisors.

A subject index entry may contain a Context Heading immediately next to the Lead Heading, to aid the searcher in judging the relevance. In DSIS, in the Context Heading, the complete subject analysis of the item is expressed (17). Context Headings to the same Lead Heading form sub-entries and are used for systematic grouping by bringing together subjects related to the Lead Heading. For this purpose of producing an 'organizing classification effect' (See note under sec. 3.5), the Context Headings in DSIS contain the indicators for the ECs and their subdivisions/divisors.

Subject index entries (other than CR entries) generally contain the address or location or bibliographical details of the source of information or document sufficient for the unique identification of it.

4.3 Cross Reference Entry

To bring together subjects related to the Lead Heading, CR entries directing the user from one Lead Heading to another are used. In DSIS, CR entries using the 'directing element' *See* are used to refer the searcher from a non-standard (synonymous or quasi-synonymous or variant form of the) heading to the standard heading. CR entries using the directing element *See also* of the ascending/descending types (See note under sec. 3.5) are not used in DSIS. The Composite Terms which are factored and taken as synonymous terms to their combination of the factored terms are given *See* CR entries in DSIS.

4.4 Permuted Cross Reference Entry

In order to provide an access through each of the significant component terms of a Complex Term, 'Permuted Cross Reference' (PCR) entries are formed in DSIS. This is done by cyclic permutation of the component terms in a Complex Term and allowing selected terms to form the Lead in the PCRs. The end of the Complex Term in such permuted or variant forms, is denoted by a slash mark, indicating the beginning of the standard form of the Complex Term. The following is a standard form of a Complex Term:

"Finishing (for) Tip Effect (using) Wax"

The following are the PCR entries, assuming that both the terms 'Tip Effect' and 'Wax' are selected to form the Lead:

"Tip Effect (using) Wax/Finishing (for)"
"Wax/Finishing (for) Tip Effect (using)"

The PCR entries do not contain any indicator digits. They do not contain any directing elements also because, the standard form of any Complex Term could be easily ascertained from the PCRs themselves. Under the standard form of the concerned PCR the other sections of the subject index entry are given.

5. Formation of Subject Headings

After formulating the name of a subject as per the DS of SIL and noting down the CR entries that are necessary, the indexer has to decide: 1) which component terms (including those forming a Complex Term) should form the Lead; and 2) which component terms should form the Context, in order to produce headings for the different subject index entries.

5.1 Selection of Lead Terms

In order to provide access, significant terms in the name of a subject are selected to form the Lead Term by prefixing a process code '$0' to the concerned term. Though it is difficult to ascertain which terms should form the Lead, certain guidelines could be followed. For instance, the term denoting the Discipline need not be selected to form the Lead, if the whole subject index is specifically for that Discipline alone. Moreover, very generic Entity terms such as Man, Plant, Animal etc., very common Property terms such as Capability, Efficiency, Cause etc., very common Action terms such as Calculation, Determination, Evaluation etc., and terms denoting Common Modifiers need not necessarily be

selected to form Lead Terms. But it should be decided by taking the name of the subject as a whole and the user community to be served into consideration. It is helpful to select each of the Complex Terms as such to form the Lead. A useful quideline for selecting Lead Terms is the Canon of Sought Heading of Ranganathan (18).

5.2 Selection of Context Terms

The Context Heading sets the context in which the Lead Heading occurs. In order to provide the maximum context, the Context Heading in DSIS in general represents the full subject analysis along with the superordinates and indicators for each of the component terms. This is helpful in creating an organizing sequence among the Context Headings to a particular Lead Heading.

If the purpose is only to serve the comprehending function then the superordinates included at the 'Modulation' step (See sec. 3.3) may be omitted, forming a 'Short Context Heading', provided it represents the full meaning of the subject. As it has been observed that "the syntactical role of each term in a heading is largely expressed by its position relative to the other terms and that in some cases, the position of a term is not of its own accord sufficient to indicate its role beyond a reasonable doubt, which means that an element of ambiguity will be present" (19), the sequence of the terms in the Context Heading is kept invariant along with the different indicator digits in DSIS. For this purpose while selecting terms to form such 'Short Context Heading', it is necessary to select the last component term (it may be a Compound Term or a Complex Term) in each of the ECs and Common Modifiers. If the selected last component term does not by itself individualise it (which happens in general, when the selected term is a Part or a Constituent), then successive superordinate terms should also be selected so that it gets individualised and is homonym free. These superordinates are "Upper Links that resolve the homonym" and the process of their selection is similar to that of the Chain Indexing system (18, p 299). Terms selected to form Context Heading in DSIS are prefixed with a process code '$1'.

5.3 Upper Link Specifiers to Lead Term

A name of a subject formulated according to the DS of SIL can be considered as a Chain having as links each of the component terms (Compound Terms and Complex Terms each taken as a unit component term) (18, p 285–8; 8, p 63). For a particular component term, all the other terms occurring prior (earlier) to it, when arranged according to the syntax rules, form Upper Links. When a term becomes the Lead Term, some of the Upper Links could be suffixed to it to further specify the Lead Term. While selecting terms for forming Context Heading, care has been taken to see that the terms selected are such that the Short Context Heading is unambiguous and homonym free (See sec. 5.2). Hence the Upper Links to a term under consideration, that are selected to form Short Context Heading have been used to form Upper Link Specifiers to the concerned term when it becomes the Lead. The sequence of component terms in the Lead Heading containing Upper Link Specifiers taken from left to right is the reverse of the sequence of the terms arranged according to the rules of syntax. This 'reverse rendering' (20) has been found necessary in retrieval (21).

5.4 Default Lead and Context

When a component term is neither selected to form Lead nor Context, the default is that it is selected to form Lead. If a component term is selected to form only Context, it is not considered for Lead at all. If a component term is the last component term of the EC manifestation, then it is selected to form the Context. This is done by checking the indicator digit with that of the succeeding term. If the difference in the indicator digits is only due to the succeeding term's indicator having a '.6' denoting Species/Type, then no action is taken. To avoid a term (neither selected to form Lead nor Context) being considered for the default options, a null process code '$9' is prefixed to the concerned term. These default options are set automatically by computer manipulation.

5.5 Processing Codes

The following are the processing codes used in DSIS for computer manipulation: 1) '$0' – Lead Term; 2) '$1' – Context Term; 3) '<' (starter), '>' (arrester) – enclosed within, is a Complex Term; 4) '$2' – Lead in PCR arising out of Complex Term; 5) '$*' (auxiliary word identifier) '/' (auxiliary word delimiter) – enclosed within, is an auxiliary/function word(s); and 6) '$9' – neither Lead nor Context.

Apart from the above process codes a special process code '$3' is used with modifiers of kind 2 to create automatically Compound Terms. For ex: "8 Skin 8.6 $3 Pig 8.6 $3 Cured" would produce "8 Skin 8.6 Pig Skin 8.6 Cured Pig Skin".

6. Coding of the Name of Subject

The formalised name of a subject 'modulated' and 'standardised' given as example in section 3.3 is:

> "Leather Technology 8 Leather 8.1 Tanning 8.1.6 Mineral Tanning 8.1.6 Chrome Tanning 8.1.6 Two Bath Chrome Tanning 8.1.5 (using) Dichromate 8.1.1 Evaluation".

In order to form the 'Short Context Heading', it is sufficient if the terms 'Leather', 'Two Bath Chrome Tanning (using) Dichromate' and 'Evaluation' are selected. As the term 'Leather' itself resolves any homonym that may arise, it is not necessary to select the Discipline term 'Leather Technology' to form the Short Context. These terms are prefixed with the process code '$1', indicating that they form (Short) Context Heading. If we assume that the subject index is only for 'Leather Technology', then it is not necessary to select 'it' to form the Lead either. Hence it is prefixed with the null process code '$9'. For the same reason, the term 'Leather' need not be selected to form Lead. However the terms 'Tanning', 'Mineral Tanning', 'Chrome Tanning' and 'Two Bath Chrome Tanning (using) Dichromate' may be selected to form the Lead. These terms are prefixed with the pro-

cess code '$0' to indicate that they are Lead Terms. The Complex Term is enclosed within angular brackets and the auxiliary/function word between '$*' and '/'. It may be necessary to form a PCR entry using the term 'Dichromate' and hence it is prefixed with the process code '$2'. The term 'Evaluation' is a very common Action term and it need not be selected to form the Lead. These decisions of the indexer are incorporated, to form the input name of a subject given below:

"$9 Leather Technology 8 $1 Leather 8.1 $0 Tanning 8.1.6 $0 Mineral Tanning 8.1.6 $0 Chrome Tanning 8.1.6 0I < Two Bath Chrome Tanning 8.1.5 $* (using) / $2 Dichromate > 8.1.1 $1 Evaluation".

The above input could be further simplified by taking default options wherever applicable and using the process code '$3' for Modifier of kind 2 to form Compound Term, as given below:

"$9 Leather Technology 8 $1 Leather 8.1 Tanning 8.1.6 $3 Mineral 8.1.6 Chrome Tanning 8.1.6 < Two Bath Chrome Tanning 8.1.5 $* (using) / $2 Dichromate > 8.1.1 $1 Evaluation".

Note: Most of the decisions relating to Lead and Context terms selection could be left to the default options available. But synonyms, quasisynonyms and synonyms due to 'factoring' of Composite Terms are to be noted separately to form CR entries to be included in the index before final sorting and printing.

7. Computer Formation of Index Headings

7.1 Initial Computer Processing

To begin with, the bibliographical record is read from a file and the coded name of a subject in it is separated out. Each of the component terms between EC indicators are pulled out along with their process codes, and the process codes are converted into flags for easier manipulation. Each component term along with its process flags is formed as individual entry in a Table built in the main memory. Each entry in this Memory Table has the following fields: 1) Lead flag (L); 2) Context falg (C); 3) Modifier flag (M); 4) Permutation flag of the constituent of Complex Term (P); 5) Component term or constituent of Complex Term (CT); 6) punctuation or type font code, if any (PN) (not used at present); 7) Starter '<' or Arrester '>' or Delimiter '/' (ASD); and 8) EC indicator of the next component term (CY).

The *L*ead, *C*ontext, *M*odifier and *P*ermutation (L, C, M, P) fields are flagged depending on the process codes for the concerned component term as per the following scheme: 1) $0 indicating Lead Term, flags L-field as 1; 2) $1, indicating Context Term, flags C-field as 1; 3) $2, indicating Lead in PCR entries flags both M-field and P-field as 2; 4) $3, indicating Modifier of kind 2, flags M-field as 3; 5) $9, indicating neither Lead nor Context, flags both L and C fields as 0; 6) $* indicating auxiliary/function word of Modifier of Kind 1, flags M-field as*; and 7) the L and C fields that are not flagged are left blank for flagging according to the default options.

The PN, ASD and CY fields are filled up with the respective data for the concerned term inserted in the CT-field. The CY-field contains the indicator for the next component term and a hash mark for the last component term. The Memory Table-1 so built for the example given in section 6 is given below:

L	C	M	P	CT	PN	ASD	CY
0	0			Leather Technology			8
	1			Leather			8.1
				Tanning			8.1.6
		3		Mineral			8.1.6
				Chrome Tanning			8.1.6
				Two Bath Chrome Tanning		<	8.1.5
		*		(Using)		/	
		2	2	Dichromate		>	8.1.1
	1			Evaluation			#

Table 1

Note: The input could be prepared in the format of the Memory Table-1 given above and input to DSIS for further processing. At present the maximum length of CT-field is 136 characters. There can be any number of Complex and Compound Terms. But the maximum number of entries in the Memory Table-1 is 50.

If Memory Table 1 contains any Modifier of kind 2 term identified by '3' in its M-field, then a Compound Term is formed prefixing it to the earlier modifyee term with certain changes in the flags. For instance, if the earlier modifyee is selected to form Context, it is nullified and the newly formed Compound Term is flagged as selected to form the Context. The flags for all the default options are then set. The Memory Table 2 so formed from the Memory Table 1 is as follows:

L	C	M	P	CT	PN	ASD	CY
0	0			Leather Technology			8
0	1			Leather			8.1
*d	1	0		Tanning			8.1.6
*d	1	0		Mineral Tanning			8.1.6
*c							
*d	1	0		Chrome Tanning			8.1.6
*d	1	1		Two Bath Chrome Tanning		<	8.1.5
		*		(Using)		/	
		2	2	Dichromate		>	8.1.1
0	1			Evaluation			#

*d = Default options set; *c = Compound Term formed.
RNO.A001

Table 2

The component terms in the Memory Table 2, are manipulated as indicated by the flags to produce different types of subject index headings.

7.2 Basic Algorithms for Manipulation

7.2.1 Formation of Complex Terms

The first constituent term of the Complex Term is identified by the presence of a starter angular bracket in its ASD field. Depending on whether Lead Heading or Context Heading is being formed, the flag in L-field or C-field is checked. Based on these flags, the first constituent term is pushed to the Lead or Context position as appropriate. Subsequent constituent terms of the Complex Term are suffixed successively in the sequence in which they occur in the Memory Table from top to

bottom till the term with an arrester angular bracket in its ASD field is also pushed. The indicator digits are omitted if a Lead Heading is being formed and are included if a Context Heading is being formed. The following is the Complex Term formed in Lead Heading position from Table 2 shown earlier:

Two Bath Chrome Tanning (Using) Dichromate

7.2.2 Formation of PCR entries

In order to form PCR entries, the earliest constituent term in a Complex Term having a '2' in its P-field is pushed to the Lead Position. The rest of the constituent terms are suffixed to it. When the last component term of the Complex Term having the 'arrester' in its ASD-field is also pushed to the Lead Heading a '/' is inserted next to it. The rest of the constituent terms occurring earlier in the Table to the term currently occupying the Lead position are suffixed successively from top to bottom of the Table. After writing out one PCR entry the next entry is formed using the next constituent term having a '2' in its P-field to occupy the Lead position and so on. The PCR entries are formed only when a Lead Heading is being formed. The only PCR entry that would be formed from Table-2 given earlier is as follows:

Dichromate / Two Bath Chrome Tanning (Using)

7.2.3 Formation of Uni-component Term Lead Heading

In order to form Lead Headings containing only a uni-component term, each of the Complex Terms as a whole and each of the other component terms which are occupants of the CT-field in the Table are considered as a unit. Those component terms flagged as '1' in the L-field are selected to form independently Lead Terms, and Lead Headings are formed for each of them. The following are the Lead Headings formed from Table-2:

Tanning	A001
Mineral Tanning	A001
Chrome Tanning	A001
Two Bath Chrome Tanning (Using) Dichromate	A001
Dichromate / Two Bath Chrome Tanning (Using)	

If Lead Headings are formed using all the component terms, disregarding the flags indicating Lead Terms, the following additional Lead Headings would be formed:

Leather Technology	A001
Leather	A001
Evaluation	A001

7.2.4 Formation of Lead Headings with Upper Link Specifiers

In order to form Lead Headings with Upper Link Specifiers, the first component term selected to form the Lead (L-field '1') from the top of the table is pushed to the Lead Term position. The term immediately occurring above the component term (Upper Link) in the Table that is selected to form the Context (C-field '1') is suffixed to the Lead Term after inserting a comma. This process of suffixing Upper Links having their C-field flagged as '1' is continued till the first component term (top of the Table) is reached. The formed Lead Heading is written out. The next Lead Heading is formed in the same way with the next component term selected to form the Lead. Whenever a Complex Term is formed as the Lead Term, its associated PCR entries are also formed. The Lead Headings that would be formed with Upper Link Specifiers from Table-2 are as follows:

Tanning, Leather	A001
Mineral Tanning, Leather	A001
Two Bath Chrome Tanning (Using) Dichromate, Leather	A001
Dichromate / Two Bath Chrome Tanning (Using)	

7.2.5 Formation of Context Headings

As mentioned earlier in section 5.2 Context Heading helps the user in comprehending the subject denoted and in predicting the relevance. Also, Context Headings to a particular Lead Heading when sorted provide an 'organizing effect' by bringing together all the Context Headings belonging to the same 'D' together, within them, all belonging to the same 'E' together and so on. If full organizing effect is required then the whole name of subject with superordinates inserted at the 'Modulation' step along with the indicators are kept as the Context Heading. Such a heading is called 'Full Context Heading'. The Full Context Heading formed from the Table 2 is:

Leather Technology 8 Leather 8.1 Tanning 8.1.6 Mineral Tanning 8.1.6 Chrome Tanning 8.1.6 Two Bath Chrome Tanning 8.1.5 (Using) Dichromate 8.1.1 Evaluation

If the purpose is just to provide an aid to the user in comprehending the denotation of the subject, then it is sufficient to form 'Short Context Heading' with just the component terms selected to form so (C-field '1') along with the indicator digits occurring in their respective CY-field. The Short Context Heading formed from Table-2 is as follows:

Leather 8.1 Two Bath Chrome Tanning 8.1.5 (Using) Dichromate 8.1.1 Evaluation

8. Types of Indexes

8.1 Types of Lead Headings

In the formation of Lead Headings two major types are possible. They are: 1) Lead Headings containing as Lead Terms only those that are selected to form Lead; and 2) Lead Headings containing all the terms (including each Complex Term as a whole) as Lead Terms irrespective of the indication of Lead selection. Within each of the above two major types, two further types of Lead Headings could be formed. They are: 1) Lead Headings containing only uni-component terms (each Complex Term being treated as a unit) as Lead Terms; and 2) Lead Headings containing Lead Terms and their applicable Upper Link Specifiers. This gives rise to four types of Lead Headings.

8.2 Types of Context Headings

In the formation of Context Headings too, two different types are possible. They are: 1) 'Full Context Heading' containing all the terms in the name of subject irrespective of the Context Term selection indication; and 2) 'Short Context Heading' containing only the terms selected to form Context Heading.

8.3 Different Types of Indexes

For each of the two types of Context Headings mentioned above, each of the four types of Lead Headings mentioned in section 8.1 could be formed. This gives rise to eight types of indexes. Another four types of indexes could be formed without any Context Headings, giving rise to a total of twelve different types of indexes. Moreover, Lead Headings with Lead Term and applicable Upper Link Specifiers could be made to form Chain Index. After formulating a Lead Heading with Upper Link Specifiers, if the remaining part of the name of subject occurring below the term currently pushed to the Lead position in Table 2 (disappearing part of the chain (22)) is kept as the Context Heading, it would resemble the PRECIS-format. Altogether about 14 different types of indexes could be formed depending on the purpose, by just manipulating the L, C, M, P-fields alone. By manipulating the contents of the CY-field it is possible to create Lead Headings having "Entity-Property", "Entity-Action" etc., combinations and their inversions, as an approximation to Kaiser's 'Systematic Index'.

The CR entries due to synonyms, quasi-synonyms and factoring are included in the generated index before sorting and printing. Appendix I is a display of the index type "All Lead Terms with Full Context Heading". Appendix II is a display of the main entries in the record number sequence.

9. Conclusion

One of the salient features of this system is that it does not require in general, any *See also* CR entries of the 'ascending' or the 'descending' types, when the Context Headings used are the Full Context Headings. Consider the Context Headings given under the Lead Term 'Beamhouse Operation' in Appendix I. All the Full Context Headings having this term are grouped, which brings together all the Context Headings having narrower terms to 'Beamhouse Operation' also, such as 'Curing', 'Salt Curing' etc. If one searches 'Beamhouse Operation' he need not be directed to search also under the narrower terms to it using *See also* entries of the descending type, unlike in other indexing systems such as PRECIS (19, p 270–1). Similarly the Full Context Headings given under the Lead Term 'Curing' in Appendix I, also have the broader term 'Beamhouse Operation'. If the searcher wants information on a broader term, he gets the term from the Full Context Heading itself. Hence there is no need for *See also* entries of the 'ascending type'. Moreover, the syntax of DS of SIL requiring the Discipline to be represented as the 'first context specifying category' eliminates the need for *See also* entries of the type 'Animals *See also* Zoology' and 'Zoology *See also* Animals', practised in PRECIS (19, p 281).

Yet another feature of this system is that, by keeping the respective Full Context Headings as a 'key' to each of the main entries (bibliographic references) and sorting them on this key, an 'organizing effect' can be produced in the sequence of the main entries. When the main entries are printed in this sequence, along with their Full Context Headings printed on top of each entry as 'Feature Headings', then the index prepared for them could be just an 'associative index' using just the Lead Headings alone.

Another salient feature of DSIS is that, the special vocabulary control tool 'Classaurus' that could be used in DSIS could be created automatically from the input prepared according to the DS of SIL. It could be kept online for referring to it to prepare CR entries to control synonyms and synonyms due to factoring of Composite Terms, and for 'modulating' and 'standardising' the names of subjects. It could also be kept up-to-date always (15, 23). Also by augmenting the input names of subjects by a different set of codes, an alphabetical thesaurus could be generated automatically (24, 25, 26).

It is possible to use the Lead Terms and their associated Full Context Headings in online information retrieval system. The system could be made to display the Full Context Headings in which the 'search term' occurs and upon choosing the relevant Full Context Headings, the bibliographical references could be retrieved and displayed. If the number of Context Headings is beyond certain limits, one or more other search terms could be input to select Context Headings having a combination of them. Such a system would have built-in vocabulary control, and searching would be quite easy. Investigation in this direction is on-going.

References

(1) Bhattacharyya, G.: Fundamentals of subject indexing languages. In: Proc. 3rd Int. Study Conf. on Classif. Research, Bombay, Jan. 6–11, 1975. Bangalore, IN: DRTC 1979. p. 86–98

(2) Bhattacharyya, G.: Some significant results of current classification research in India, Int. Forum Inform. Doc. 6 (1981) No. 1, p. 11

(3) Spang-Hanssen, H.: Are classification systems similar to natural languages? In: Proc. 3rd Int. Study Conf. on Classif. Research, Bombay Jan. 6–11, 1975. Bangalore, IN: DRTC 1979. p. 15

(4) Gardin, J.-C.: Document analysis and linguistic theory. J. Doc. 29 (1973) No. 2, p. 146–147

(5) Bhattacharyya, G.: Foreword to Fugmann, R.: The analytico-synthetic foundation for large indexing and information retrieval systems. Bangalore, IN: Sarada Ranganathan Endowment for Library Science 1983. p. IX

(6) Bhattacharyya, G.: POPSI: Its fundamentals and procedure based on a general theory of subject indexing languages. Libr. Sci. Slant Doc. 16 (1979) No. 1, p. 14–15

(7) Bhattacharyya, G.: A general theory of subject indexing language. Dharwad, IN: Karnatak University, Ph. D. Thesis 1980.

(8) Ranganathan, S.R.: Prolegomena to library classification. Bombay, IN: Asia Publ. House 1967. 3rd ed. p. 422–424

(9) Bhattacharyya, G.: POPSI: a source language for organizing and associative classification. Libr. Sci. Slant Doc. 19 (1982) p. 249–252

(10) Kaiser, J.: Systematic indexing. Aslib report of proceedings. Oxford, GB, Sept. 24–27, 1926. p. 20–33. Reprinted in: OIding, R.K. (Ed.): Readings in library cataloguing. Canberra, AU: F.W. Cheshire Ltd. 1966. Reprinted again: New Delhi, IN: Lakshmi Book House 1967. p. 154

(11) Ranganathan, S.R.: Prolegomena to library classification. London, GB: Edward Goldston Ltd. 1937. p. 137

(12) Ranganathan, S.R.: Library classification: fundamentals and procedure. London, GB: Edward Goldston Ltd. 1944. p, 39

(13) Ranganathan, S.R.: Elements of library classification. Bombay, IN: Asia Publishing House 1962. p. 130

(14) Bhattacharyya, G.: Classaurus: its fundamentals design and use. In: Proc. 4th Int. Study Conf. on Classif. Research, Augsburg, DE, June 28–July 2, 1982. Frankfurt, DE: Indeks Verl. 1982. p. 139–148
(15) Devadason, F.J., Kothanda Ramanujam, M.: Computer aided construction of "alphabetic" classaurus. In: Proc. 4th Int. Study Conf. on Classif. Research, Augsburg, DE, June 28–July 2, 1982. Frankfurt, DE: Indeks Verl. 1972. p. 173–182
(16) Keen, E.M.: On the generation and searching of entries in printed subject indexes. J. Doc. 33 (1977) No. 1, p. 19
(17) Svenonius, E.: Indexical contexts. In: Proc. 4th Int. Study Conf. on Classif. Research, Augsburg, DE, June 28–July 2, 1982. Frankfurt, DE: Indeks Verl. 1982. p. 129
(18) Ranganathan, S.R.: Classified catalogue code with additional rules for dictionary catalogue code. Bombay, IN: Asia Publ. House 1964. 3rd ed. p. 44
(19) Austin, D.: PRECIS: a manual for concept analysis and subject indexing. London, GB: The Council of British National Bibliography 1974. p. 45
(20) Ranganathan, S.R.: Subject heading and facet analysis. J. Doc. 20 (1964) No. 2, p. 114
(21) Fugmann, R., Winter, J.H.: Reverse retrieval: towards analogy inferences by mechanized classification. Int. Classif. 6 (1979) No. 2, p. 85–91
(22) Mineur, B.W.: Relations in chains. J. of Libr. ship. 5 (1973) No. 3, p. 179
(23) Devadason, F.J.: Online construction of alphabetic classaurus: a vocabulary control and indexing tool. Inform. Process & Managem. 21 (1985) No. 1, p. 11–26,20 refs.
(24) Devadason, F.J., Balasubramanian, V.: Computer generation of thesaurus from structured subject propositions. Inform. Process & Managem. 17 (1981) No. 1, p. 1–11
(25) Devadason, F.J.: Postulate-based Permuted Subject Indexing Language as a metalanguage for computer-aided generation of information retrieval thesaurus. Int. Forum on Inform. & Doc. 8 (1983) No. 1, p. 22–29
(26) Devadason, F.J.: Computer based systems for generating different types of subject indexes and alphabetical classaurus based on "Deep Structure" of Subject Indexing Languages. Dharwad, In: Karnatak University, Ph. D. Thesis. (Guide: Kumbhar, M.R.) 1984. 649 p.

Appendix I

AGING RESISTANCE = OZONE RESISTANCE
AIR (WITH) CATALYST METAL SALT / OXIDATION (BY)
BEAMHOUSE OPERATION
LEATHER TECHNOLOGY 8 HIDE AND SKIN 8.4 SKIN 8.1
BEAMHOUSE OPERATION 8.1.4 CURING 8.1.2 EFFECTIVENESS 8.1.2.1 EVALUATION 8.1.5 (USING) MICROSCOPIC ANALYSIS A004
LEATHER TECHNOLOGY 8 HIDE AND SKIN 8.4 SKIN 8.6 PIG SKIN 8.1 BEAMHOUSE OPERATION 8.1.4 CURING 8.1.6 SALT CURING 8.1.6 DRY SALT CURING 8.1.5 (USING) DRUM A001
BLUING = CURING
CATALYST METAL SALT / OXIDATION (BY) AIR (WITH)
CHEMICAL PROPERTY
LEATHER TECHNOLOGY 8 LEATHER 8.2 PROPERTY 8.2.6 CHEMICAL PROPERTY 8.2.6 HYDROPHOBICITY 8.2.5 (INFLUENCED BY) ORGANO SILICON COMPOUND A 003
LEATHER TECHNOLOGY 8 LEATHER 8.2 PROPERTY 8.2.6 CHEMICAL PROPERTY 8.2.6 HYDROPHOBICITY 8.2.5 (INFLUENCED BY) TANNING A 006
CURING
LEATHER TECHNOLOGY 8 HIDE AND SKIN 8.4 SKIN 8.1 BEAMHOUSE OPERATION 8.1.4 CURING 8.1.2 EFFECTIVENESS 8.1.2.1 EVALUATION 8.1.5 (USING) MICROSCOPIC ANALYSIS A004
LEATHER TECHNOLOGY 8 HIDE AND SKIN 8.4 SKIN 8.6 PIG SKIN 8.1 BEAMHOUSE OPERATION 8.1.4 CURING 8.1.5 SALT CURING 8.1.6 DRY SALT CURING 8.1.5 (USING) DRUM A001
DETERMINATION (USING) SPECTRO PHOTOMETRY
LEATHER TECHNOLOGY 8 LEATHER CHEMICAL 8.6 SOAKING MATERIAL 8.6 SOAK LIQUOR 8.2 PROTEIN CONTENT 8.2.1 DETERMINATION 8.1.5 (USING) SPECTROPHOTOMETRY A005
DRUM / DRY SALT CURING (USING)
DRY SALT CURING (USING) DRUM
LEATHER TECHNOLOGY 8 HIDE AND SKIN 8.4 SKIN 8.6 PIG SKIN 8.1 BEAMHOUSE OPERATION 8.1.4 CURING 8.1.6 SALT CURING 8.1.6 DRY SALT CURING 8.1.5 (USING) DRUM A001
Dharam.

Appendix II

A001 RADKEVITCH, D.P., KARNET, N.S.
Preservation of pigskin with dry salt in drums. Myasanaya Industrya. 5; 1982; 13–15. Russian.
A002 KATRICH, V.W.N., KEDRIN, E.A., SURABJAN, K.M.
Influence of tanning methods on the physico mechanical and tanning properties of sole leather. Leder. 30; 5; 1979, May; 68–72. German.
A003 WAKHRAMEJEWA, V.N.V., SURABJAN, K.M.
The influence of organo silicon compounds on hide and leather as a basis for the hydrophobicity of leather. Leder. 30; 5; 1979, May; 75–6. German.
A004 ZUBIN, A.N., KOROBEJNIKOVA, LN., STEPANOVA, EH.I.
Effect of the method of preservation on the skin. Kozhen Obuv Promyshl. 21, 9; 1979; 28–31. Russian.
A005 STRAKHOV, I.P., BULGAKOVA, I.V., KOLARKOVA, A.S.
Determination of proteins in the soak liquor using spectro photometry. Kozhev Obuv Promyshl. 21, 8; 1979; 53–4. Russian.

Gesellschaft für Klassifikation eV – SIG/IS –

EK-03 (en)
August 1985

FREE TEXT IN INFORMATION SYSTEMS
Capabilities and Limitations

0. Preliminary Remarks

The German Society for Classification has started to publish Recommendations in order to help the newcomers to the field in the application of those methods that have been tested by experts since many years and have found the approval of the profession.

These Recommendations are not meant to serve as standards, as the activities of standardization are the matter of the respective German Institute for Standardization. But they could well be regarded as topics and sources for a future consideration of a kind of standard if such a necessity should be felt by the authorities.

The present Recommendation is the first one to be published also in English. (The German version appeared in Int. Classif. 1985–1, p. 23–26). Furthermore, it should be pointed out that this English version has been improved in some sections, due to responses received after publication. Any comments from readers of this recommendation are welcome to the address given below.

1. Introduction: Free text and indexing language based information systems

Computer technology has opened up new possibilities for the design and use of information systems. Once, it was, for instance, necessary to represent all documents in an indexing language*[10] employing notations, just to enable their contents to be entered into a mechanized search file. This necessitiy has been obviated by *computer facilities*, and the possibility is now often considered of foregoing the use of *any* indexing langugage and storing documents entirely or partially in their original language. Such an approach, however, ignores the fact that the indexing[9] of documents greatly facilitates their accurate[1] retrieval from a mechanized search file and that often indexing is a prerequisite for such retrieval. *This is true regardless of which technical aids and computer programs are implemented or used.* Occasionally, information systems have been placed in operation that are exclusively based on the free text method, while at the same time other systems that work with an indexing language and are based on expert indexing have been abandoned or rejected right from the start. It was hoped that in doing so costs could be saved and that the performance of expert indexing could be approximated or even exceeded.

Such a result, however, can be obtained only under very special simplified conditions. Although in its *initial stage* a free text information system may appear quite successful. As a rule, the consequences of a decision in favour of the free text method do not become obvious until a far advanced stage has been reached. The decision will be recognized to have been mistaken so late that it can hardly be corrected.

Free text information systems suffer from a steady decline of search accuracy[1]. Their *operational use* becomes more and more expensive, due to the large proportion of irrelevant responses that have to be weeded out again and again. Concomitantly, the proportion of omissions in the responses to a search continuously increases. Then the savings achieved on the input side must increasingly be paid back. It will also become increasingly necessary to restrict search activities to keep them from becoming intolerably expensive. Such an information system largely fails to achieve the purpose for which it was conceived. Often the only solution is to begin again from the start. If, at that time, no different approach is pursued, it would be impossible to achieve the goal of an operational, coherent information system in a subject field of interest.

It is true that various forms of *automatic indexing* can be used with some degree of success in retrieving documents that are stored in free text form. But on sober reflection there is at present no indication that this approach could ever achieve the *quality of expert indexing* based on a powerful and reliably employed indexing language[10]. Experiments that appear to prove the contrary have been performed under unrealistic and artificial conditions. In the appendix references to some articles are listed which concern the topic of the capabilities of controlled, human expert indexing (2–21).

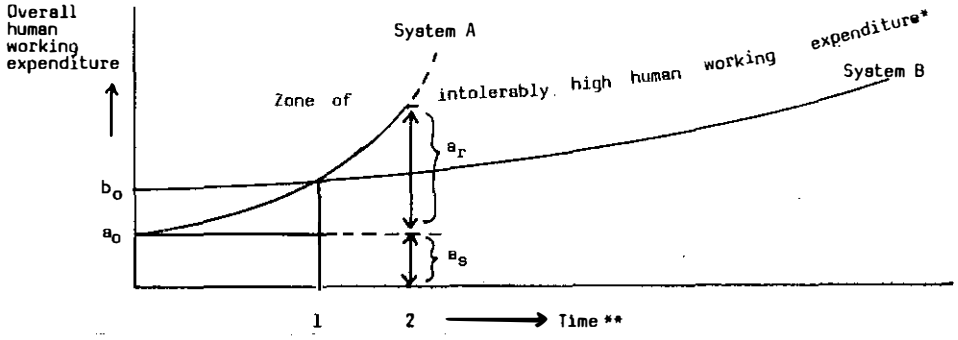
2. Input versus search costs of information systems

Two information systems A and B are depicted in Figure 1. A may be a free-text based system, i.e. a system with low initial human working expenditure a_0, because at the time the system is initialized no search costs are incurred and only relatively low input costs arise. But this system has a steep cost increase when it begins to be used and, hence, has a low survival power. By contrast, information system B, which may typically be an indexing language based system, is much more expensive in its initial stage, but its overall operation costs, which include the search and which are largely determined by the amount of human work involved, will be lower beyond the break even point 1. From time 2 onwards it is the only surviving system, provided it is given the chance to get through its costly start-up period. Hence, the costs for system B must be seen as the minimum investment required to achieve the envi-

* The meaning in which these terms are used in this text is explained in the appended glossary.

Herausgeber: Gesellschaft für Klassifikation eV. Sekretariat: Woogstr. 36a, D-6000 Frankfurt 50. Bearbeiter: Spezielle Interessengruppe Indexierungssprachen (SIG-IS). Nachdrucke erhältlich beim Sekretariat.

Dependence of human working expenditure in a information system on time



a_r: Human working expenditure for retrieval a_s: Human working expenditure for storage

* The tolerable human working expendure is assumed to increase slightly with the size of the data base. This is due to its increasing usefulness.

** Date base size and search frequency are assumed to increase concomitantly with time.

sioned objective of a powerful and durable information system.

When one proceeds from full text storage to the storage of abstracts of texts and, finally to the assignment of freely chosen keywords to documents to be entered into the system, one concomitantly approximates the type of information systems that are based on an indexing language. But the latter are still markedly different from the free text systems in their performance and behaviour so that it is justified to separate them conceptually. Below we will briefly discuss the pecularities of free text systems in comparison with systems based on a indexing language. We shall investigate how these pecularities affect the accuracy[1] of retrieval and the requirements placed on the expertise of the systems' operators. These effects are most apparent in the prototype of the free text[8] method which uses full text storage.

An information system can be used for various supplemental applications other than to search for concepts[3], if it also has available the full text of documents. For example, texts located in a search can be made to appear on a terminal screen, they can serve for terminological studies, when, for example, the various meanings that are associated in the literature with a particular mode of expression are being investigated. But these capabilities of full text systems are not dealt with in this paper.

3. Loss of relevant[15] responses

General concepts[4] constitute the main problem in free text data bases because they can be expressed in many different ways in natural language. For an adequate search all conceivable modes of expression would have to be compiled as (alternative) search parameters. This would have to include not only synonyms, but also all conceivable varieties of phrases and other nonlexical[12] expressions for the concept under consideration. Since this is inherently impossible, queries for these searches are bound to be incomplete, and necessarily the responses to such queries will be correspondingly incomplete. The kind and size of the incompleteness depend on chance and are not apparent from the results of the search. It has therefore often been ignored in the evaluation of an information system. Accordingly, the free text method is justifiable only when in searches for general concepts information[11] loss to an indeterminate extent is acceptable.

If in a uniform manner the authors use standard expressions for a concept, perhaps even exclusively those of the lexical[12] kind, then these expressions are better predictable, and loss in information retrieval can be kept within limits or even largely suppressed. This is particularly true for *individual concepts*[5] (cf. 2,4,21). As a rule, it is not difficult to memorize or to look up all modes of expressions that have been in use for an individual concept and to list them exhaustively as alternative search parameters, especially when searches are performed only rarely.

In natural language, the *hierarchical relations* that prevail among concepts in a viewpoint of interest, are only rarely (and only incompletely) displayed. If we want to include under a general search concept (e.g. insects or fungi) all the subconcepts pertaining to it (e.g. ants, locusts, mosquitoes etc.), an intolerably large number of search words would have to be compiled for searching a free text data base. Searches of this kind in such data bases necessarily entail large information losses.

In some types of text the natural language mode of expression is subject to certain *restrictions*, which improves their predictability at the time of the search. Medical and experimental reports complying with a prescribed format are examples. Here, the risk that information will be lost is correspondingly reduced.

If input into a data base is carried out for an extended time or if documents are written in different languages, a particularly large lingual heterogeneity and unpredictability develops in the data base.

This heterogeneity becomes increasingly more difficult to grasp in its entirety. The result is increasingly incomplete queries and a continually growing loss of information.

He who *searches in a book* using its table of contents can be expected to be more familiar with the modes of expression used in the book, and for the preparation of indexes of this kind free text methods are more appropriate than in data bases exhibiting a larger lingual heterogeneity.

In an indexing language based retrieval system information loss can also occur. This happens when an important concept is overlooked by the indexers. Such losses can be countered, however, by making it obligatory for the indexers to select concepts consistently from a set of prescribed semantic categories[2], provided that these concepts do not occur in an entirely extraneous context.

Occasionally, the concepts of an inquiry and the meaning of expressions in a document are so poorly defined that great uncertainty accompanies their translation into an indexing language. Then it is useful if, at least in addition to representing these concepts in an indexing language, the possibility is provided for searching for the wording as presented in the original text. For this purpose an *additional free text search capability* can be of great service.

The extent to which the responses to a query are incomplete may not be detected for a long time and even then revealed only through large efforts. Loss of relevant information is therefore often underestimated or even ignored in many evaluation experiments.

4. Irrelevant responses

Responses that are irrelevant to an inquiry can occur in free text systems, when in the retrieved documents the search words appear with an undesired meaning and/or an unexpected context. This will hardly occur in indexing language based systems, because here ambiguity is resolved, and the context in which a concept is embedded can be represented in a predictable form. The broader the range of subjects in a data base and the longer the texts of the documents, the larger will also be the meaning divergence and incorrect contextualization. For this reason abstracts are often preferable to full texts in data bases.

Incorrect contextualization cannot satisfactorily be overcome by requiring the search words to appear in more or less close proximity, because the proximity in which the search words co-occur in the original texts can often not be anticipated. – Consequently, the inclusion of proximity criteria in a query often leads to loss of relevant[15] information. The satisfaction often sensed in the use of proximity parameters would quickly vanish if the searchers became aware of the extent of information loss incurred by the use of these parameters.

With data bases that are (and remain) small a relatively high proportion of noise can be tolerated. But this is not true of large and/or fast growing files. In the latter, the large *absolute amount of noise* can develop into a critical obstacle (cf. diagram above). Noise that has developed to an intolerable level during the mere maintenance and updating of the data base can lead to the break down of the system and, hence, eventually to the *complete loss of all the information stored in its files.*

In an indexing language based system *noise can be kept under control* by designing an indexing language that is suitably effective and by assuring that it is employed in a sufficiently reliable manner. It must, however, not be tailored merely to the *present* needs. Much more, it must also provide accuracy reserves to meet the certainly higher demands of the future. In particular, an indexing language should, in addition to its vocabulary, possess a sufficiently well developed *grammar.* This will also be conducive to indexing reliability, because the vocabulary of these indexing languages can be kept small and, thus, easy to memorize and overview. The development of grammar and its usage in a *predictable* manner is an investment that must be accepted for the sake of the survival power of an information system.

It is true that among irrelevant responses to an inquiry a searcher will occasionally find some that he will rate as interesting (pertinent[13]) ones. But it cannot be the task of an information system designed to retrieve documents of a *predefined* kind to submit documents that do not meet the requirements of the search.

5. Staff Qualification Requirements

Free text systems appear markedly user-friendly, because they oviate the necessitiy of learning a special indexing language. This has, however, by no means made these systems attractive to the casual user to the expected extent. To use these systems effectively and economically, *information specialists are still required*, as has been proved in practice (cf. e.g. 22). At present, there is no indication that this state of affairs will change in the foreseeable future. All statements to the contrary can only be true for very simple inquiries, for example for individual concepts[5]. But even here, the casual inquirer must put up with a *loss of information*, the kind and extent of which will often go unnoticed by him.

Furthermore, user-friendliness becomes questionable when it requires the user to guess and try out a tremendous number of conceivable modes of expression for his search concepts, especially when he cannot tell when and if he has reached an end in his trials.

An argument advanced in favor of free-text systems is that they do not require the specialized knowledge in information science that is necessary for the design and maintenance of indexing language based systems. But here one must accept the inevitable consequence of the renunciation of expert knowledge, namely a correspondingly *low level of performance.* It might be noted, however, that in the design and maintenance of indexing languages, too, this expert knowledge is sometimes missing, which also adversely affects the performance of the retrieval systems employing them.

A free text system may, however, serve a useful purpose in *collecting* those concepts that should be represented in an indexing language. At the same time it makes the texts containing them available for search, if only in an improvised fashion. Thus, a free text system may well constitute a useful preliminary stage in the development of an indexing language based system. In this function it can be retained even after the indexing language has been put into operation. Concepts that are not yet represented in the vocabulary can, thus, immediately be accessed and kept available for the later decision whether or not they should be included in the vocabulary.

Another advantage sometimes claimed for free text systems is that the input into them can be carried out by persons who have no specialized knowledge of the subject field dealt with in the texts. But it would be more correct to say that *a lack of specialized knowledge* does not become so quickly obvious here as when one is working with an indexing language. During the continual translation of the essence of the texts into an indexing language faulty understanding will make itself immediately apparent in various ways, so that remedies can be applied early to prevent complications from developing later.

6. Conclusion

Indexing languages are used in information systems not merely because they were once indispensible for storage

purposes, but *because our natural, technical and colloquial languages are usually poorly suited to the task of retrieving stored texts.* This is due to the *lack of predictability of their modes of expression* for technical concepts, especially the general ones.

Although free text systems can effectively complement indexing language based information systems, the former can rarely replace the latter altogether, at least if the system is to continue to perform in an adaquate manner beyond its initial development stage.

The most widespread and oldest style of leadership is "management by ignorance". The occasionally disastrous consequences of this style have been sufficiently understood since the outcome of the Trojan War. That so astonishingly little effort is being made to avoid its consequences in the information field is explained by the circumstance that too many people in the information field have too little knowledge of classification theory (which must not be confused with computer science) and that the advice of the experts in this field is given no credence.

The recommendations made here are intended to call attention to the existence of this knowledge and to encourage its use.

Glossary of terms

1 *Accuracy*: The accuracy of the information supply is determined by the extent to which the loss of relevant[15] responses and the noise of irrelevant responses is suppressed (2).

2 *Category*: General concept[4], with respect to which no more general concept existis which is meaningful from the perspective of the experts of the subject field (1,2).

3 *Concept:* The entirety of the true and essential statements that can be made about the referent[14] (1).

4 *Concept, general*: Concept[3] with respect to which at least one more specific concept exists which is meaningful from the perspective of the experts in the field. – It includes only relatively few conceptual features[7] (2).

5 *Concept, individual*: Concept[3] with respect to which no more specific concept exists which is meaningful from the perspective of the experts in the subject field (2).

6 *Controlled vocabulary*: Vocabulary which contains those descriptors which are permitted for indexing.

7 *Feature of a referent*[14]: Component of the definition of a concept, being expressed through one true and essential statement about the referent to which the concept appertains.

8 *Free text method*: Words, parts of sentences or sentences in natural, technical or colloquial language are used for the storage and retrieval of texts in mechanized information[11] systems.

9 *Indexing*: The process of discerning the essence of a document and representing this essence with a sufficient degree of predictability and fidelity (usually in an indexing language) (2).

10 *Indexing language*: Language which represents concepts[3] and statements of a document with a sufficient degree of predictability and fidelity. – The natural language mode of expression of an author usually lacks predictability of the mode of expression, especially in the case of general concepts[4] and statements (2).

11 *Information*: Any message that proves to be of interest to a recipient (2).

12 *Lexical unit*: Linear string or character, specifically agreed upon to denote a concept[3]. – A unique location can always be assigned to a lexical unit where it can be entered or looked up in an alphanumerical arrangement.

13 *Pertinence*: Any message that proves to be of interest, irrespective of whether it satisfies the parameters of a query, ranks as pertinent and hence, as information[11]. – Accordingly, even irrelevant responses may turn out to be pertinent. – An inquirer is often unable, to define his entire field of interest by means of concepts[3] and concept relations. He can therefore not expect an information system to supply him with all those messages in the system that he would find interesting if he encountered them and no other messages but those (2).

14 *Referent*: Anything about which statements can be made (1).

15 *Relevance*: A message is considered relevant if it comprises all the concepts[3] and concept relations (as far as they can both be defined in the inquiry) in the desired or in a higher degree of specificity.

16 *Subject heading*: Lexical unit[12] to represent a concept with great conciseness, frequently chosen as a replacement for a paraphrasing, less concise mode of expression for a concept.

References

(1) Dahlberg, I.: Grundlagen universaler Wissensordnung. München, DE: K.G. Saur Verl. 1974. p. 7–10

(2) Fugmann, R.: Role of theory in chemical information systems. J. Chem. Inform. Comput. Sci. 22 (1982) p.119

(3) Fugmann, R.: Natural vs. indexing language in chemical documentation. Angew. Chemie Int. Ed. Engl. 21 (1982) p. 608–616

(4) Fugmann, R.: The complementarity of natural and indexing languages. Int. Classif. 9 (1982) p. 140–144

(5) König, E.: Verbindliche versus freie Indexierung. In: Numerische und nicht-numerische Klassifikation zwischen Theorie und Praxis. Frankfurt/M., DE: Indeks-Verl. 1982, p.263–270. = Studien zur Klassifikation, 10

(6) Rolling, L.: EDV-unterstützte Manipulierung von Begriffsbeziehungen. Nachr. Dokum. 24 (1973) p. 60–64

(7) Hersey, D.F., Foster, W.F., Stalder, E.W., Carlson, W.T.: Free text word retrieval and scientist indexing. J. Doc. 27 (1971) p. 167–183

(8) Henzler, R.G.: Free or controlled vocabularies. Int. Classif. 5 (1978) p. 21

(9) Duckitt, P.: The value of controlled indexing in full text data bases. 5th Online Meeting, London, GB 1981.

(10) Rothman, J.: Online searching and paperless information. J. Amer. Soc. Inform. Sci. 32 (1981) p. 77

(11) Svenonius, E.: Natural language vs. controlled vocabulary. Proc. 14th Can. Conf. on Inform. Sci. 1976.

(12) Schwarz, Chr.: Freitext-Recherche, Möglichkeiten und Grenzen. Nachr. Dokum. 33 (1982) p. 228

(13) Gebhardt, F., Stellmacher, I.: Design criteria for documentation retrieval languages. J. Amer. Soc. Inform. Sci. 29 (1978) p. 191–199

(14) Soergel, D.: Indexing languages and thesauri. Los Angeles, US: Melville 1974, pp. 29, 57

(15) Wellisch, H.: The cybernetics of bibliographic control: Toward a theory of document retrieval systems. J. Amer. Soc. Inform. Sci. 31 (1980) p. 41

(16) Yerkey, A.N.: A preserved context indexing system for microcomputers: PERMDEX. Inform. Process. Management 19 (1983) p. 165–171

(17) Wall, R.A.: Intelligent indexing and retrieval: A man-machine-partnership. Inform. Process. Management 16 (1980) p. 73–90

(18) Moses, P.B., Nelson, L.E.: Indexing and abstracting chemical information: The view of two industrial chemists. J. Chem. Inform. Comput. Sci. 24 (1984) p. 189–190

(19) Fugmann, R.: Zum Finanzierungsproblem bei der "Ware" Information. Mittbl. Nr. 4 der Fachgr. Chemie-Information, Frankfurt/M, DE: Ges. Dt. Chemiker 1983. p. 3–6

(20) Blair, D.C., Maron, M.E.: An evaluation of retrieval effectiveness for a full-text document-retrieval system. Comm. ACM 28 (1985) p.289–299

(21) Fugmann, R.: The five-axiom theory of indexing and information supply. J. Amer. Soc. Inform. Sci. 36 (1985) p. 116–129

(22) Krentz, D.A.: On-line searching – specialist required. J. Chem. Inform. Comput. Sci. 18 (1978) p. 4–9

Published by: Gesellschaft für Klassifikation eV. Secretariat: Woogstr.36a, D-6000 Frankfurt 50. Editors: Special Interest Group Indexing Languages (SIG-IS). Reprints available from the Secretariat.