## Bibliographical notes:

The lexicon discussed in the present study and the theoretical framework (the 'text-structure world-structure theory' /TeSWeST/), of which it is a component are analysed from different points of view in the following studies:

van Dijk, T. A., Ihwe, J., Petöfi, J. S., Rieser, H.: *Zur Bestimmung narrativer Strukturen auf der Grundlage von Textgrammatiken*. (= Papiere zur Textlinguistik, Band 1 und 1A). Hamburg: Buske 1972. (2. Auflage mit einem Nachwort von H. Rieser, 1974).

Petöfi, J. S.: *Towards an empirically motivated grammatical theory of verbal texts*. In: Petöfi, J. S., H. Rieser (Eds.): Studies in Text Grammar. Dordrecht: Reidel 1973.

Petöfi, J. S.: *Zum Aufbau eines „Lexikons"*. In: Rave, D., Brinckmann, H., Grimmer, K. (Eds.): Syntax und Semantik juristischer Texte. Darmstadt 1972.

Petöfi, J. S.: *Modalität und topic-comment in einer logisch fundierten Textgrammatik*. In: Dahl, O. (Ed.): Topic and comment, contextual boundness and focus. (= Papiere zur Textlinguistik, Band 6). Hamburg: Buske 1974.

Petöfi, J. S., Rieser, H.: *Präsuppositionen und Folgerungen in der Textgrammatik*. In: Petöfi, J. S., Franck, D. (Eds.): Präsuppositionen in Philosophie und Linguistik/ Presuppositions in Philosophy and Linguistics. Frankfurt: Athenäum 1973.

Petöfi, J. S., Rieser, H.: *Probleme der modelltheoretischen Interpretation von Texten*. (= Papiere zur Textlinguistik, Band 7) Hamburg: Buske 1974.

It should be noted that information gained from documentation theory and thesaurus research has influenced the construction of the TeSWeST. Cf.:

Petöfi, J. S.: *A tezaurusz-kérdés jelenlegi helyzete*. [The present state of the thesaurus problem]. Budapest: OMKDK 1969.

Petöfi, J. S.: *On the problems of co-textual analysis of texts* COLING (International Conference on Computational Linguistics in Stockholm), Preprint No. 50. 1969. In German translation in: Ihwe, J. (Ed.): Literaturwissenschaft und Linguistik I–III. Frankfurt: Athenäum 1972.

Regarding the present state of the empirical work aiming at the construction of a multi-purpose thesaurus/lexicon cf. the description of the research-project "Aufbau eines axiomatischen Kern-Lexikons der deutschen Sprache" in this issue (p. 99).

Robert Fugmann
Hoechst AG, Frankfurt/M.-Höchst

# The Glamour and the Misery of the Thesaurus Approach

## Treatise IV on Information Retrieval Theory[1]

Fugmann, R.: **The Glamour and the Misery of the Thesaurus Approach**.
In: Intern. Classificat. 1 (1974) No. 2, p. 76–86

If any important natural-language term which a documentalist encounters in storing literature and in phrasing enquiries is admitted as an addition to a thesaurus, then the thesaurus will soon exceed the limits of its operancy and will increasingly fail to serve the purpose of an efficient device for reliable terminological control in the input and retrieval stage. This continous decline can effectively be counteracted by conceptual analysis of candidate terms and by resynthesis of the terms of their conceptual constituents. This suggests a balanced combination of the thesaurus and the analytico-synthetic classification approach, particularly in large information retrieval systems. The representation of certain, predominantly syntactical relations, however, exceeds the capabilities of both approaches. These relations can be managed by two different devices described, namely by a clearly defined set of relation indicators and by an optionally additional graphical representation of extended concept relations. (Author)

### 1. Introduction

Any mechanized literature search aims at retrieving documents from a file that are relevant to the special topic of the inquirer. In order to enable the search mechanism to perform this task the inquirer will have to *define* the special goal of his literature search. In such a search request it must be laid down *in advance*, i. e. without any previous knowledge about relevant documents contained in the file, which particular features should be possessed by the desired documents and are to be considered as an indication of their relevance to the special topic of the inquirer (cf. 1, postulate of definability, p. 134). This is at least true of a test search directed to a sample of the entire file, on the basis of which the request can be modified and then directed to the entire file. In particular, it must be laid down in advance in the

[1] Extended version of a paper presented at the Third International Conference on Classification Research, Bombay, January 1975 First treatise: Ref. 1; Second treatise: Ref. 3; Third treatise: Ref. 15

request (and, depending on the kind of search strategy chosen, sometimes also in programmed machine instructions) which particular words, word connections or other strings of characters are required to occur in the desired documents, representing the conceptual features that are of interest to the inquirer. The search mechanism has performed its task to perfection if on the one hand all documents that meet these a priori established search requirements and are contained in the file have been retrieved and if, on the other hand, no documents have been retrieved that do not (or not sufficiently exactly) satisfy these requirements.

In spite of the perfection of the search mechanism an inquirer might miss interesting documents or encounter "non-interesting" ones among those retrieved. Then an attempt must be made to depict the goal of the literature search with a higher degree of fidelity, in the request if that is possible in the particular language of the retrieval system in which the request must be phrased.

Any request of whatever kind can lead to the desired result only if, among other things, it is free from requirements that are *unsatisfiable* by relevant documents in the file(2, 3, 4). An inquirer always runs the risk of stating unsatisfiable requirements if he approaches a mechanized file in the same manner as he is accustomed to do in his conventional, self-service literature searches, a typical feature of which is that they are based less on difinition than on intuition. Furthermore, he will rarely be aware of the many input conventions which are obeyed by the indexers and the knowledge of which is essential for successful retrieval.

In particular, an inquirer is likely to state search requirements that are too trivial or too specific for the file or are unsatisfiable merely by reason of their linguistic presentation.

As far as search requirements are concerned that are too trivial in nature, the chemist as indexer will refrain from filing a kind of knowledge with which he has been familiar since the time of his education ("Chlorine belongs to the group of halogen elements", "Treating alcohols with acetic anhydrid leads to acetic acid esters" etc.). Hardly any chemist will desire directions of this kind from a mechanised file since they would constitute no genuine information for him. Entering them into the file would raise input costs considerably and at the same time seriously depreciate the file as a source of valuable information, i. e. of items of knowledge that may directly or indirectly influence the decisions of the searcher ((5), (6)) or remove uncertainty (7) on his part.

On the other hand a mechanised file, e. g. for chemical literature, should not be expected to provide highly specific information in other fields of knowledge. For instance, in the context of highly resistant lubricants an author might have dealt in detail with the geometrical shape of the cogs of gear wheels. A chemist as an indexer, however, should not be expected to analyze and file these details, e. g. the mathematical functions that describe the surface shape of the cogs. This would most probably exceed his professional competence and furthermore prove futile, for hardly any expert interested in these details would consult a file for *chemical* literature. For, he would encounter a far too narrow coverage of his literature and would badly miss professionally competent assistance in phrasing his search request.

A most common source of search failure that we shall investigate in this article at some length originates from search requirements that aim at *a particular mode of expression* for the topic of an inquirer, although it is normally entirely immaterial for the inquirer, at least in all the fields of natural science and technology, in which way the topic he is interested in was expressed in relevant documents, provided they are presented in a language that he can understand. If modes of expression different from that expressly laid down in the search request have always been entered into the file, and if the search mechanism was not instructed on the conceptual equivalence of all these different modes of expression *in advance*, then loss of relevant information is bound to occur. Avoiding unsatisfiable search requirements of this kind necessitates the *predictability* (1 (postulate of predictability), 8) of the modes of expression by which the topic of an inquirer has been expressed in the file, i. e. the predictability of the representation of *concepts and concept relations.*

In order to achieve this predictability one may keep a continuous record on which particular modes of expression for topics of possible enquiries have already been entered into the file, perhaps because the authors of relevant documents have used these expressions. Since this vocabulary can be established and updated *after* the corresponding documents were filed, it can be denoted an "a-posteriori-vocabulary", irrespective of the kind of language (natural or not) of the vocabulary terms.

From this point of view the more conventional kind of vocabulary, such as classification schedules, controlled vocabularies, authority lists etc. constitutes an "a-priori-vocabulary". The natural-language representations for essential concepts and concept relations in a document to be filed are translated into one single, solely permissible mode of expression *before* they are entered into the search file and are thus made predictable. Such vocabularies have also to be established *before* the corresponding documents are entered into the search file. In phrasing a search request, only a single or at most a very few related modes of expression taken from this vocabulary need be considered.

For natural-language vocabularies of index languages, in particular for those also desplaying relations among the individual terms, the denotation *"Thesaurus"* has become common, irrespective of whether or not they are used as a-priori or a-posteriori vocabularies. On this meaning of the word "thesaurus" the following treatise is based.

Thesauri (cf. 13) hold a number of significant advantages as compared with conventional classifications. The latter require the assignment of notations to all concepts of present *and future interest* to enable the indexer to analyse all incoming documents without delay and without previous revision of the vocabulary. However, such an a-priori-vocabulary can always only be established incompletely, in particular as far as newly emerging concepts that are continuously being created by research and development are concerned.

Furthermore, many classifications provide only limited hospitality for concepts and concept relations newly emerging or having only recently deserved interest.

Thus, in the course of the years these classifications are bound to become obsolete or chaotic as a result of illogical insertions and extensions. A thesaurus, however, provides an apparently unlimited hospitality for new concepts and their relations.

By virtue of the abovementioned features a thesaurus can, at least in part, be used as an a-posteriori vocabulary and thus relieve the documentalist of the time-consuming, expensive, and always incomplete preparatory work for classification schedules. A thesaurus also offers the possibility of displaying relations other than hierarchical ones, should this be of interest, and of dealing with terms the meaning of which is obscure. A thesaurus does not necessitate problematic decisions with respect to the meaning of a term as is the case with classifications of the conventional kind.

When using a thesaurus one is, due to the natural language character of its terms, not compelled to look up almost all of them as is necessary in the case of classifications. – One should, however, beware of overestimating this particular feature of a thesaurus, for, if looking up is entirely dispensed with, this may also inpair the consistency of indexing depth and, concomitantly, the predictability of the representation of concepts and concept relations and thus devaluate the thesaurus for the purpose for which it is intended. Furthermore, reflecting about (instead of looking up) the most appropriate, concise, and sufficiently common natural language mode of expression for a topic to be represented may frequently outweigh the timesavings gained by omitting look-up. Replacing a mode of expression for a concept consisting of several words or sentences by an appropriate concise term is strongly in the interest of the reliability of the search.

Due to the aformentioned favourable features of a thesaurus it has often been considered superior to classifications of the conventional kind, particularly with respect to the criterion of survival power under the constraints of the future requirements and to the ease of implementation of a thesaurus-based retrieval system and of its operational utilization.

As a vocabulary of natural language terms, however, a thesaurus necessarily suffers from several deficiencies inherent to natural language if it is used for information retrieval purposes. Among these are the ambiguity of many terms and their vacillation in meaning over the years. For many important and quite common concepts no concise terms have been coined as yet, a situation which necessitates the use of entire sentences for expressing them or at least highly multiworded terms. Both kind of expressions are however little suited information retrieval purposes.

Furthermore, a natural language term only rarely expresses all the essential features of an object in the viewpoint of a certain subject field. Therefore, natural language terms of even closely related concepts have most frequently no string of characters (word stem, syllable etc.) in common that could display genuine relationships and might be suited for phrasing a generic search as is possible in the case of systematic notations. Generic searching with natural language terms therefore requires a (sometimes prohibitively large) number of alternative search terms. In addition, natural language terms in mechanized information retrieval may simulate concepts that are not implied (see chapter 3). All these deficiencies of natural language in retrieval systems have already been thoroughly investigated, in particular by Ranganathan and his Indian school (cf 9).

The ease with which newly emerging terms and relations between terms can, purely physically, be entered into a thesaurus, has seduced documentalists into making excessive use of this possibility. This has sometimes led to an entire break down of prominent and initially promising thesauri after several years of operational use or, at the very least, to a continuous decrease in their usefulness for the purpose for which they have originally been intended, namely to attain and to maintain the predictability of the modes of expression for topics in a search file. We shall investigate in some detail the nature of this pernicious and sometimes even fatal process. It will become apparent that this process can effectively be counteracted by recalling to mind *principles typical of analytico-synthetic classification and by employing them in a balanced and proper combination together with the thesaurus approach*.

## 2. The demands made on thesauri for large, operational retrieval systems

If an indexer is to translate a topic of a document or an inquiry into a predictable mode of expression with the aid of a thesaurus, then he must not content himself with selecting from the thesaurus *some* acceptable term which he encounters or memorizes more or less contingently. Rather is it paramount for the effectiveness of the subsequent retrieval process that he selects the *most approporiate mode of expression* for this topic that is provided by the thesaurus (1, p. 132, postulate of fidelity). Classification and indexing has always been based on this fundamental principle.

Only in this way can the intersubjective and intrasubjective consistency and, hence, predictability of the translation into the index language be attained, and a retrieval system of an advanced precision and recall be achieved.

A document may, among many other things, deal in detail with countermeasures against a particular kind of destruction of copper materials, characterized by the occurrence of many sharply localized centers of chemical demolition of pipes caused by impurity particles contained in drinking water. If *any* multi-or single-word term of practical importance representing a more or less composite concept related to the phenomenon of corrosion prevention is admitted to the thesaurus, then the lattter may provide, possibly among others, the following terms for the topic of this document:
Antirust compositions
Cavitation prevention
Corrosion inhibitors for metals
Corrosion prevention in stainless steel vessels
Prevention of undesired chemical change of materials
Pitting prevention

Spalling prevention
Stabilizing against chemical destruction
Stabilizing of copper against corrosion.

A thesaurus to which composite terms of the kind listed above are admitted will necessarily be (or soon become!) very large, comprehending perhaps several tenthousands of terms.

From the above preselection of thesaurus terms an indexer may, after some reflection, choose the terms "pitting prevention" and/or "Stabilizing copper against corrosion", which he regards the most appropriate ones among those preselected and, hopefully, among all those available in the thesaurus. A still more appropriate term is "pitting prevention in copper pipes (for drinking water)". It was, however, not contained in the thesaurus for good reasons or the indexer happened to overlook it.

The reliability with which the most appropriate term will be chosen also depends on the extent to which alien terms are undesirably included in this preselection. For, these may render it unnecessarily large and lead to considerable scattering of the terms most appropriate for the representation of a topic. Thus, they may depreciate it for the purpose for which it is intended, namely to present an offer of terms easy to survey for the purpose of choosing the most appropriate one. From this point of view, a situation extremely disserviceable to the indexer is the one shown in figure 1 A (see below). In this figure a broken line means a term in a thesaurus alien to a topic under consideration. The related terms are expressed by full lines, the thickness of which is intended to indicate the closeness of their relation to the topic.

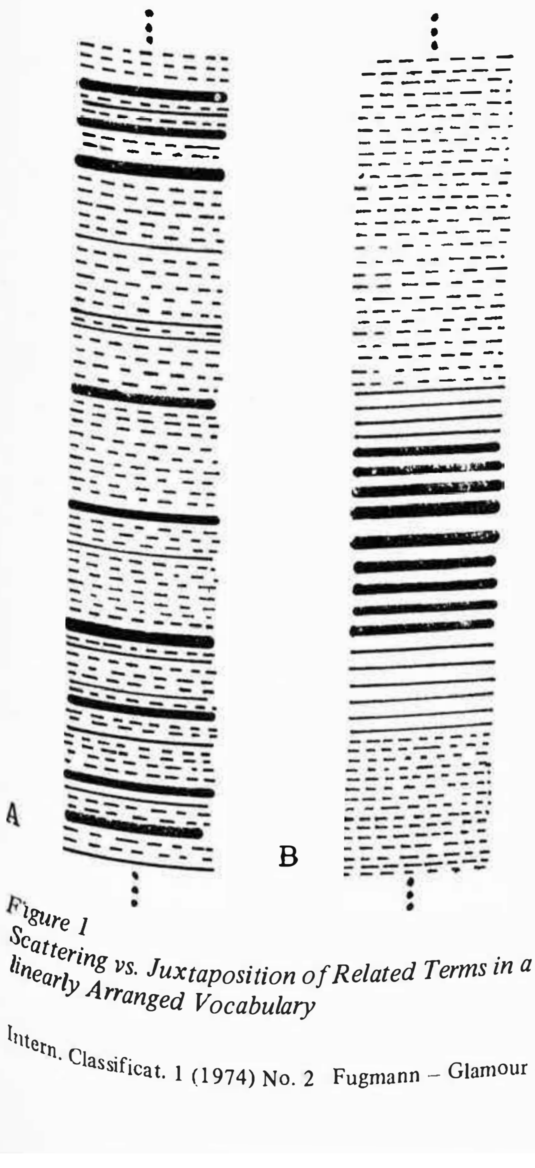

*Figure 1*
*Scattering vs. Juxtaposition of Related Terms in a linearly Arranged Vocabulary*

Considerable expenditure with respect to patience, concentration, and memory will be demanded from the indexer, if he is expected to survey large preselections of this kind (for example several hundred related terms) originating from very large vocabularies and to add to it by purely memorizing terms that are still missing in it. These resources of patience, concentration and memory, however, are in each case subject to natural human limitations, and the indexer is permanently in danger of being overtaxed in their employment, particularly in systems that are continually growing with respect to file size, vocabulary size and number of topics to deal with in storage and/or retrieval every day. When this kind of overburdening occurs, the indexer will fail intersubjectively and intrasubjectively to assign consistently to the documents and search topics the most appropriate terms provided by the thesaurus (1, p. 126, postulate of available search capacity).

This will impair the predictability of the modes of expression for concepts and concept relations. Correspondingly defective will be the searches based on such an indexing work. They will suffer from ballast of irrelevant information and from loss of relevant information contained in the file.

Much more useful for the indexer is the situation demonstrated in figure 1 B. Here, the indexer encounters an orderly arrangement of terms in *juxtaposition* (10), more or less closely related to the topic under consideration, brought together at a foresseeable place, most ideally according to Ranganathans APUPA pattern (9). This will strongly facilitate the reliable and fast choice of the most appropriate thesaurus term even under the pressure of operational work.

The larger a vocabulary and the larger the number of topics to deal with in every day storage and retrieval practice the less will an indexer be able to compensate for lacking order of the kind shown in figure 1 A by employment of his limited resources or time, patience, and memory. The greater will also have to be the importance attached to an orderly arrangement of the kind shown in figure 1 B.

In practice, however, it is only possible to approximate to the ideal state of the extreme high degree of order of figure 1 B. For, the multidimensionality of relations that prevails by nature among the terms cannot undistortedly be depicted in a strictly one-dimensional arrangement of term lists, as we know from our experience with conventional classifications and with prototypes of thesauri such as Roget's.

In a strictly linear sequence of terms, relationship can be expressed only to a very limited extent by arranging them in juxtaposition. For, in such a sequence, a certain term (e. g. no. 100) can have only one single predecessor or successor (e. g. no. 99 or no. 101) with which the term no. 100 is related in one particular viewpoint (e. g. viewpoint $\alpha$). Those terms that are related with this term in a *different* viewpoint $\beta$, e. g. the terms nos. 121–130, 216, 375 and 810, cannot be brought in juxtaposition to term no. 100, since these positions are already occupied by the terms no. 99 and no. 101. The same holds for terms that may be related with term no. 100 in a third or fourth viewpoint. Many of them will be scattered over the linearly arranged vocabularly. They can be linked with term no. 100 at most by references.

In a two-dimensional display of thesaurus terms on the other hand, it is possible to arrange *several* terms in juxtaposition to a certain, central term such as no. 100. Thus, here it is possible, at least to a certain extent, to display relationship among terms in more than one single, privileged viewpoint (11, 12).

Highly composite terms of the kind for which "Pitting in copper pipes for drinking water" is an extreme example, are if unrestrictedly admitted to an unprotected thesaurus, not only the most common source of its unlimited growth but will also lead inherently, even if more latently, to a correspondingly *large network of term relations in the thesaurus*. The majority of these "relations" will be purely syntactical in nature and of the kind object – process or object – property or process – process conditions, if literature in the field of natural science is covered. Such relations are unlimited in variety. For example, in a thesaurus of the kind described above the terms "copper pipes for drinking water", and "copper pipes" and "drinking water" and "copper" and "pipes" can also be expected, as well as terms such as "copper prices", "copper purification by electrolysis", "copper ore exploration", "influencing electrical conductivity of copper" etc. Thousands of such terms will have to be linked in some way with each other in the course of several years' large-scale operational use of a thesaurus. For, the indexer will have to be safely directed from the thesaurus term "copper" or "drinking water" to terms that are more appropriate to represent the topic of a document or of an inquiry, e. g. to "copper pipes for drinking water".

It is the strength of a thesaurus as compared with conventional classifications that almost any kind of relationship, including the hierarchical one, can in principle be expressed by permanent juxtaposition of related terms or by reference notes. By pursuing the network of relations marked by these reference notes one can also *generate* juxtaposition for related terms, either purely in the mind, or with the aid of paper and pencil, or using a computer print-out or a cathodic ray display screen. If, however, no effective safeguard was established against the unlimited, continual afflux of composite terms that experts' terminology continually produces in the course of the years then an *extremely heavily branched and extended network of relations of the most varied kinds will also develop during the practical use of a thesaurus.* An ever increasing expenditure in pursuing such a network of relations will be involved in generating the desired, most helpful kind of juxtaposition of all thesaurus terms to be considered for the representation of a topic. Whatever mechanical means are employed in this process (if any) nothing relieves the indexer from scanning and scrutinizing such a preselection of terms in order to choose the most appropriate one. This is particularly true of unprotected thesauri if they are applied to a large field of literature coverage or even on the global scale.

The situation is still further aggravated if in the references to "related" terms it is not specified in *which particular viewpoint* relationship between the terms prevails. One may, e. g. be interested in being directed from "copper" to "copper pipes" or "containers" or "reaction vessels", and from here to the "use of copper materials for the conduction of water, especially of drinking water", but not from "copper" to "copper conductivity", "copper prices", "copper purification" etc. and to hundreds of other "related" terms. However, specifying the particular viewpoint of term relationship, which is a typical classificatory principle, would require identification and limitation of the kind of relationships to be recognized among terms, i. e. a technique not common in our present day thesaurus approach.

It must be concluded that in large information retrieval systems some limitation must be imposed on the size and growth of a vocabulary and on the kind of term relations recognized in it. In other words *not every important term or relation can be admitted to a thesaurus*, because this would, in the course of time, inevitably lead to its overexpansion and to that of the network of relations prevailing in it. This would entail its entire breakdown or at least a continuous decrease in its effectiveness.

How can this statement be reconciled with the requirement for large systems to represent *any* topic of present and future interest with the utmost degree of fidelity? Obviously, those topics the terms for which would have to be rejected as additions to a thesaurus, must not entirely be neglected but represented instead in a way different from that of introducing a new thesaurus term for them. They will have to be represented by postcoordination of terms already available or justifyable in a thesaurus (*analysis*) *and* preserving the particular kind of connectivity of the individual components isolated in such an analysis (*synthesis*).

Both processes will have to be performed in a predictable way, i. e. not based on intuition and on the pragmatics of a present situation and on the contingencies of the present day natural language expression of a concept, but much more on a priori established principles and postulates that concentrate on the conceptual content of a term.

Hence, in large retrieval systems (or in those that will soon become large), admittance of terms to a thesaurus should be restricted to those terms that represent meaningful conceptual constituents such as the isolates in an analytico-synthetic classification.

## 3. Analysis

If it is agreed that an inquirer is primarily interested in documents related to his search *idea* and not so much in the occurrence in relevant documents of *words* that he happened to choose to express his idea, then it is obvious that it is the *concept* that has to be analysed into its constituents and not the word or word sequence with which the idea happened to be expressed in a document or in a inquiry. Only on the basis of such a concept-oriented analysis is it possible to avoid the isolation of misleading concepts or of those that are meaningless in isolation (e. g. "silver" from "silver fir", "German silver", "silver thaw", "silver jubilee"; "high" from "high temperature", "high altitude", "high yield"; "hand" from "on the other hand" etc.). Only in this way can concepts be made explicit that would otherwise remain hidden, for example "lung" in "tuberculosis" (infestation of the lung by mycobacterium tuberculosis), "pneumonia" etc. In other words, the analysis will have to be based on a (hopefully generally accepted) definition of the term.

In order to render this analysis predictable it has to be established a priori which categories of conceptual constituents should be isolated in this analysis as elementary constituents and should, then, resist further analysis. This is a procedure the principles of which are known from semantic factoring, analytico-synthetic classification and several other variants of conceptual analysis.

In the IDC[1] system for chemistry and related fields the following set of semantic categories has proved successful:

matter
living entity (and organs)
apparatus
process (chemical, nonchemical)
property and state of matter.

It is advisable to exclude from this conceptual analysis only those concepts that would yield highly ubiquitous conceptual constituents such as water, air etc., when contained in "water solubility", "steam engine", "water tight", i. a., concepts that would hardly be phrased as search requirements in isolation and would, if kept unanalysed, not entail an unacceptably high expansion of the vocabulary. If they were consistently analysed too, they would cause considerable expenditure as a result of the need to employ synthetic devices that would have to display the connectivity between "water" and "solubility" etc. (cf 14, p. 72).

The analysis can also be omitted (or made only optional to the indexer) in the case of such terms in which nothing like an agreed-upon definition exists. The natural language term as such is then entered into the vocabulary and can be phrased as a search requirement. It must then be left to the inquirer whether or not the meaning of this natural language term, as implied by an author, fits into what the inquirer understands by this term. It is the strength of a system that combines the thesaurus with the classification approach that it can master these terms, the analysis of which would be futile and an object of permanent controversy among the documentalists on the one hand and between the documentalists and the users of the system on the other.

Another large group of terms should not be permitted to claim much of the valuable space in a thesaurus: terms for concepts such as persons (or generally living entities), institutions, countries (or generally geographical terms such as rivers, mountains etc.). It is typical of these concepts that no reasonable, more specific concept can be derived from them by adding a conceptual feature. Thus, these concepts are essentially not generic in nature. Often they have not even a generic concept above them, at least none that is of practical importance in a retrieval system for the particular field of interest. Hence, in the case of terms for these concepts, there is no requirement for a offer of related terms to facilitate the choice of the most appropriate one. They need therefore not be arranged in a group of related terms and they need not be contained in a thesaurus that is intended to serve just this purpose. If the predictability of the mode of expression for such concepts is not assured per se, as is the case with the names of persons, institutions, countries etc., it is sufficient to list them in a purely formal, alphabetical arrangement in order to govern synonymy. They are therefore better taken care of by a dictionary (the purpose of which has often been confused with the different one of a thesaurus).

Should, even after these restricting measures, a thesaurus still embrace more than, say, 3000 terms, then the reliable application of these terms is still endangered. It would require too much patience, time, attention, and memory to survey it reliably, at least within a reasonable time span of familiarization with it. If an extreme reliability in the assignment of thesaurus terms is paramount, as is the case in the patent field, where searches with practically 100 % recall are highly desirable, then it is advisable to segregate from a large thesaurus a vocabulary with a maximum of 2–3000 particularly important, privileged terms. The employment of terms from the rest of the vocabulary in storage and retrieval is not excluded in such a approach. But they will have to be used with due caution in phrasing a search in order to avoid loss of relevant information. A discussion of the properties and various possibilities of such a vocabulary of privileged terms is, however, not the subject of this paper.

## 4. Synthesis

Merely analysing composite concepts into their conceptual constituents, without representing the particular kind of relation that prevailed among these constituents previously, would not constitute a sufficiently accurate representation of a composite concept and would violate the postulate of fidelity. From such an unstructured conglomerate of conceptual constituents any misleading combination could be read out in a mechanised search. For example, if in a document several chemical substances were described, each with a set of particular properties of its own, it is necessary to preserve the linkages between an individual substance and its particular properties on the one hand and also between the various reported properties of an individual substance on the other. An inacceptably low precision of searches for substances and their properties would otherwise result. This is especially true of large systems and in particular of systems of a high indexing exhaustivity. Thus, raising the indexing exhaustivity, intended as a means of increasing the quality of searches in a retrieval system, has often lead to a break down of this retrieval system after having been in operational use for some time, if no sufficiently effective synthetic devices were employed.

Therefore, in any advanced indexing method the analysis of composite concepts is succeeded by a (re–) synthesis of this concept. This synthesis is intended to lead to a predictable representation of the relations between the

[1] IDC: International Documentation in Chemistry, 6 Frankfurt am Main, Hamburger Allee 26, Federal Republic of Germany

conceptual constituents isolated by the analysis. It is the predictability of this representation that distinguishes the re--synthesised concept from its unanalysed, natural-language representation.

## 5. Synthesis with graphical methods

The most perfect way of representing concept connections is the graphical one. In this way almost any kind of concept relation can be represented in a predictable way suitable for mechanised searches. The IDC uses the TOSAR* system for part of the documents of its literature coverage (15). To each process recorded in a document a set of three levels is assigned in which, respectively, the initial situation, the process itself, and the results of a process is represented by nodes. These are occupied by the corresponding concepts. The individual nodes are connected by edges. The outcome of a process can be the initial situation for another process succeeding in time etc.

For example, in the following TOSAR graph a sequence of processes is represented in which, as a first step, an iron surface is treated with trichloroethylene as a solvent and 20 % aqueous phosphoric acid at 20° C for 15 minutes. This results in degreasing and, simultaneously, etching of the surface. This surface is posttreated with 5 % aqueous phosphoric acid at 50° C for 5 minutes. The surface prepared in this way is coated with a layer of either an alloy of iron and manganese or an alloy of iron, chromium and nickel or, finally, with a layer of a polyamide. The surface is thus effectively protected against corrosion:

* TOSAR: Topological Representation of Synthetic and Analytical Relations of Concepts.

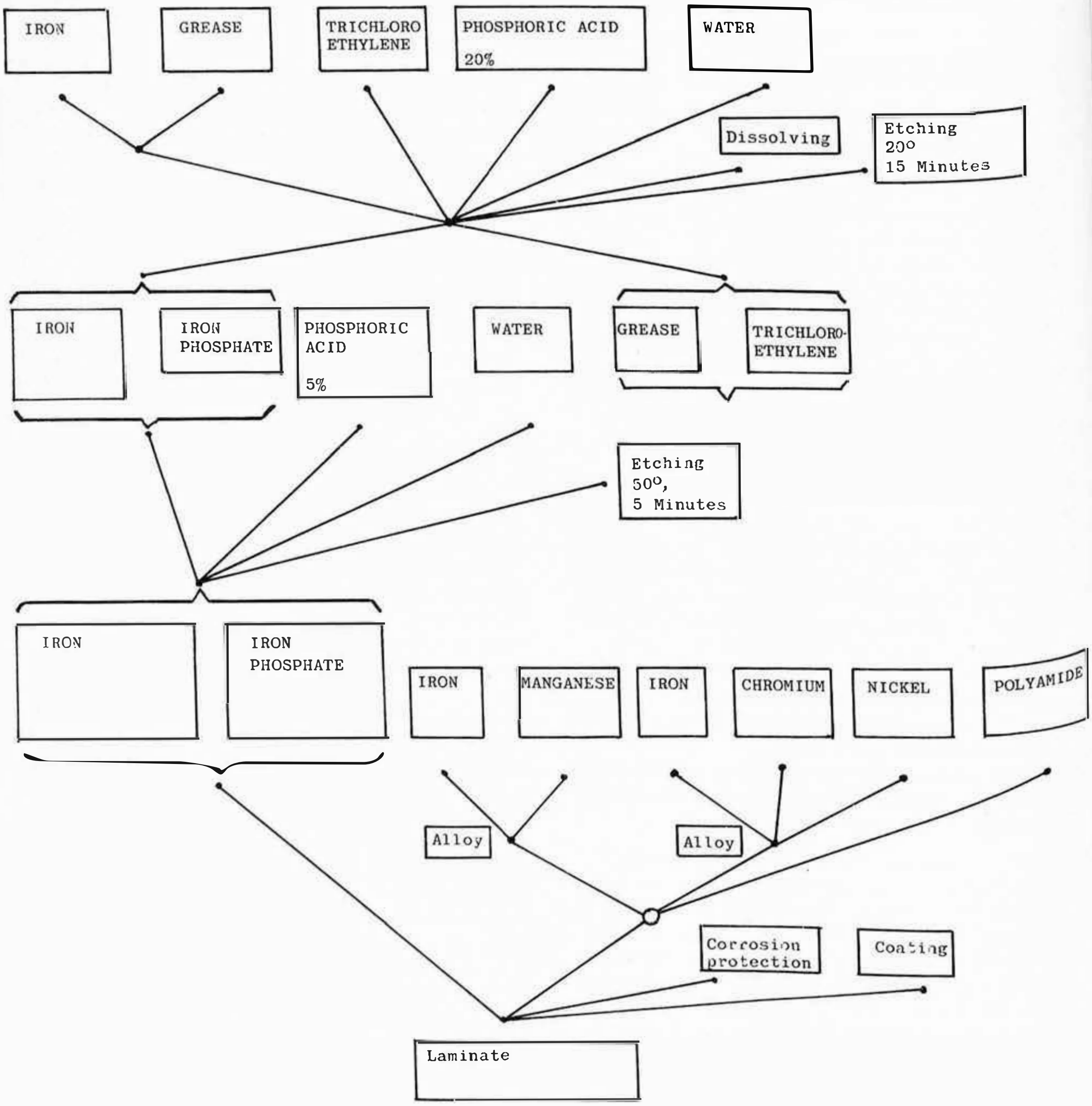

*Figure 2*
*Graphical Representation of a Sequence of Processes in a TOSAR Graph*

We are presently investigating whether the lucid syntactical structure of the TOSAR graph is also suitable as the backbone of an *absolute syntax* for expressing any kind of scientific statement. In particular, each triplett of levels comprises subject, predicate, objects and attributes of a complete sentence. Expressing scientific statements in such an elementary, standardized format would considerably facilitate their proper mechanised analysis and processing for storage and retrieval.

## 6. Synthesis by means of relation indicators

The graphical representation of the above figure is not yet sufficient if high demands are made on the precision of the retrieval and, correspondingly, on the representation in the file. For example, the participants in the coating process should be differentiated as to which of them is the *substrate* to be coated (iron) and which is the *coating materials* (alloys or polymers). Otherwise, in a mechanised search, documents of the kind described in the above figure might be confused with those in which, conversely, the polymer or an alloy material as a substrate is coated.

In order to express the particular kind of participation of an object in a process, a special kind of relation indicators similar to those of Diemer and Henrichs (16) was developed in IDC. For each process (for example "Dissolving", "Coating") a family of relation indicators exists, for example:

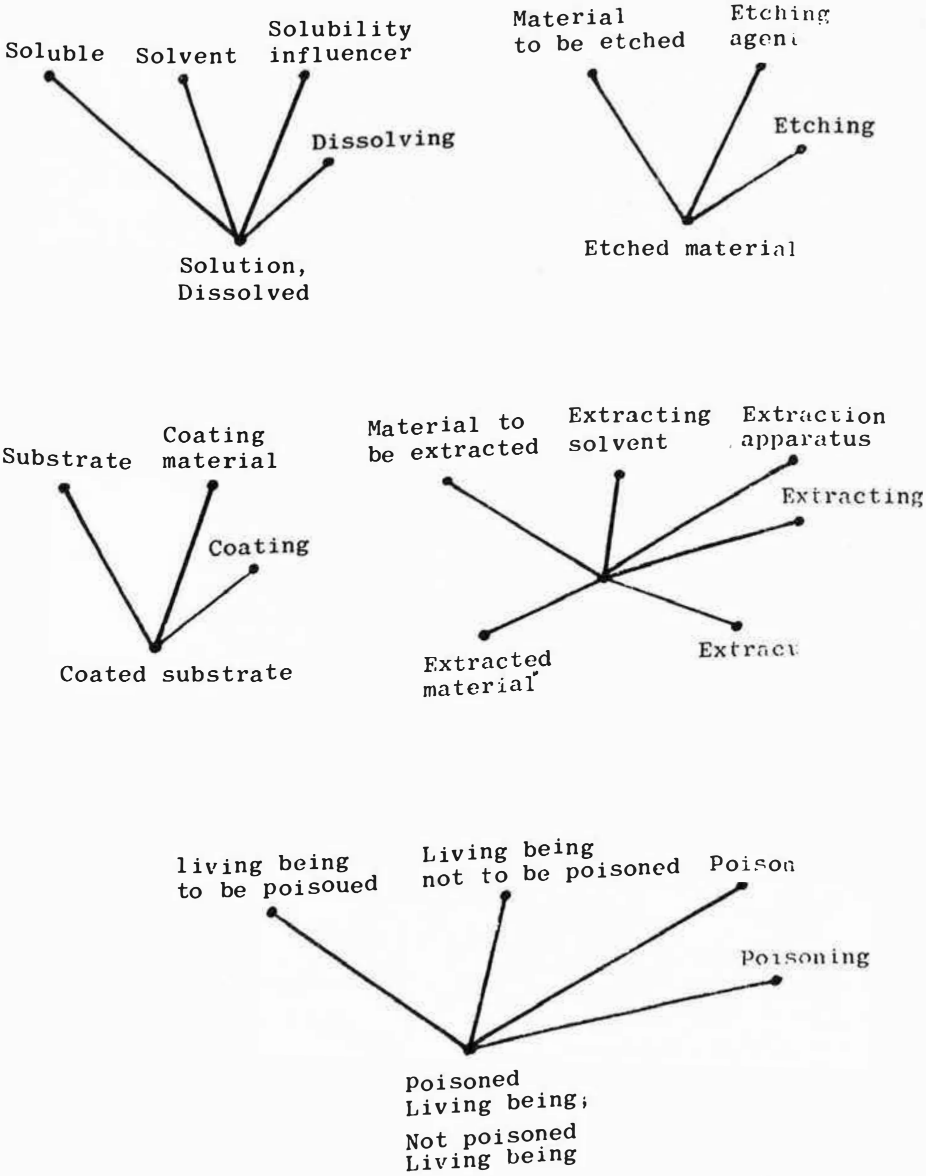

*Figure 3*
*Graphical Representation of some Families of Relation Indicators*

If for a process a typical apparatus or typical accelerators or inhibitors exist, then these also belong to the family of relation indicators. For reasons not to be discussed in detail in this paper, systematic notations were assigned to these relation indicators in the IDC system. These notations lucidly display both their appurtenance to the same family of relation indicators as well as their position in a hierarchy of generically related families.

For example, the notation of the family "dissolving" is closely related to that of the family "mixing", "dispersing", and "emulsifying". These notations also permit the systematic formation of negatives, for example to represent negated concepts such as "insoluble", "not subject to poisoning" etc.

Thus, the relation indicator very specifically reflects the *role* played by a certain object (substance, living entity, apparatus, energy etc.) in the context of an inquiry or a document. The consistent assignment of the appropriate relation indicator to any object involved in a process at the same time links to one another all those objects that participate in some respect or other in one and the same process. For, in the relation indicators assigned to these objects the string of characters recurring is that which has been made typical of the concept family of this process in the particular index language. An example is the string "sol" in the family consisting of "soluble", "solvent", "solubility influencer", "dissolving", "solution", "dissolved", "insoluble" etc., if natural language terms are used or the string "BEA" if for certain reasons documentary language terms of the kind "7BEABX", "7BEAGX", "7BEAHX", "6BEAXX", "8BEAMX", "8BEANX", "8BEABU" were used. Phrasing such a string of characters as a search requirement in conjunction with the

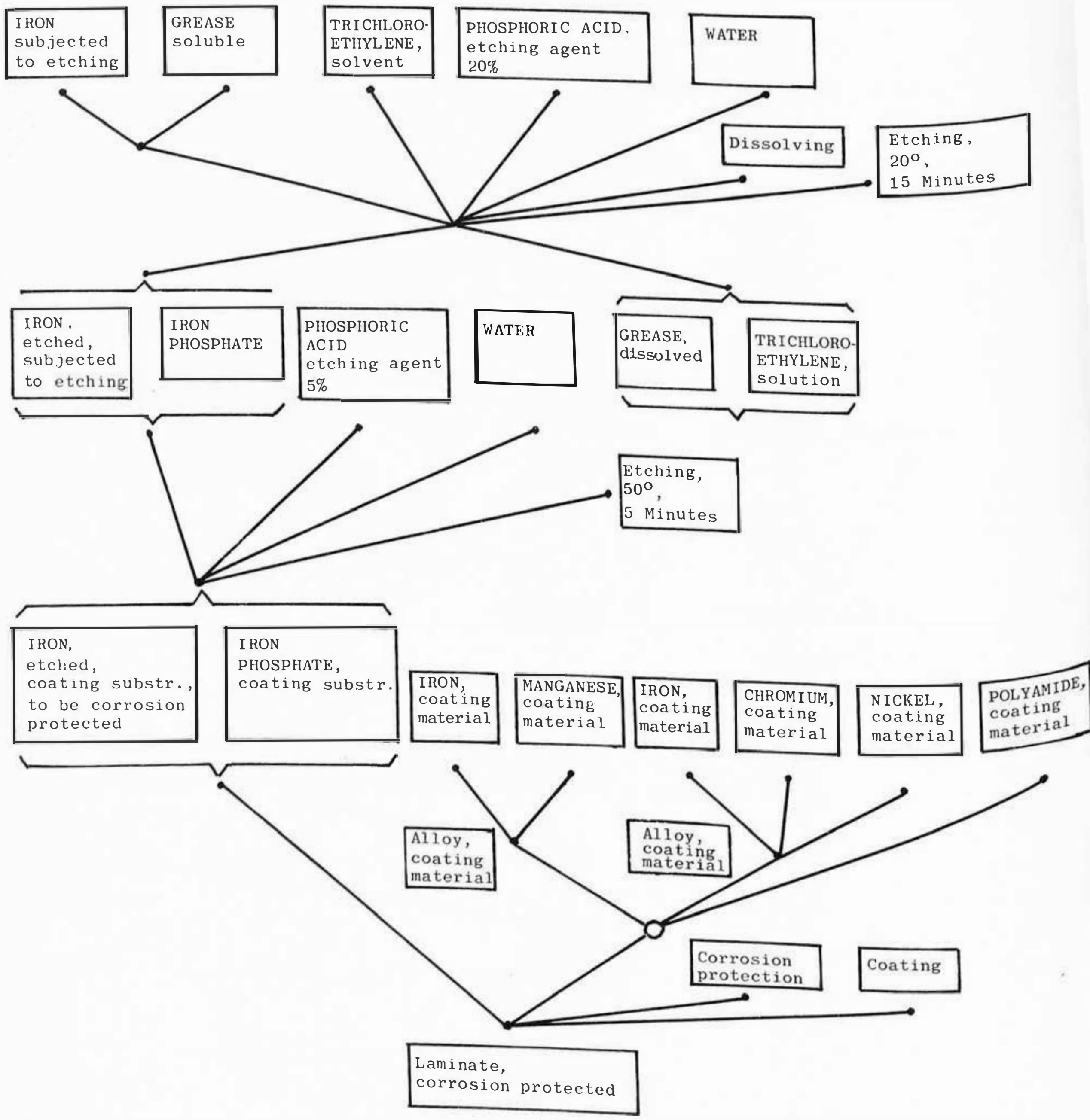
IRON subjected to etching
GREASE soluble
TRICHLORO-ETHYLENE, solvent
PHOSPHORIC ACID, etching agent 20%
WATER
Dissolving
Etching, 20°, 15 Minutes
IRON, etched, subjected to etching
IRON PHOSPHATE
PHOSPHORIC ACID etching agent 5%
WATER
GREASE, dissolved
TRICHLORO-ETHYLENE, solution
Etching, 50°, 5 Minutes
IRON, etched, coating substr., to be corrosion protected
IRON PHOSPHATE, coating substr.
IRON, coating material
MANGANESE, coating material
IRON, coating material
CHROMIUM, coating material
NICKEL, coating material
POLYAMIDE, coating material
Alloy, coating material
Alloy, coating material
Corrosion protection
Coating
Laminate, corrosion protected

*Figure 4*
*TOSAR Graph for a Sequence of Processes Including Relation Indicators*

corresponding objects avoids confusing those objects with one another that were in fact recorded in a document but were – undesirably – involved in different processes and were therefore recorded in a context and in a function different from that which an inquirer may have in mind. – The relation indicators will have to be assigned to the corresponding objects even in those cases in which the particular function of an object was not expressly stated by an author. Therefore, the consistency of their usage by the indexers is relatively high, which is conducive to a correspondingly high recall ratio in the search responses.

This method of using relation indicators as search requirements is a procedure quite similar to that of employing *links* in phrasing a request. Thus, *the relation indicator in a sense combines in itself the functions of conventional role indicators and of links*. The relation indicator is markedly different from the latter ones in that it is generally applicable (it can in principle be derived from any process concept) and in that it is fairly well defined. Roles and links of the conventional kind have often been devoid of these features. This deficiency has often seriously impaired their consistent usage and has sometimes caused authors generally to question the value of roles and links in retrieval practice.

If one superimposes the graph for the processes with the graphs for the individual families of relation indicators, the complete graph ready for encoding and storing results, as shown in Figure 4.

Each box indicates an individual file record in which closely related concepts are brought together. The connection of the concepts of each individual record with those of other records is preserved in the form of an encoded graph on a separate magnetic tape ("syntactic tape", cf 15).

In particular, a record of a chemical substance (or substance group) comprises all its properties (among which the relation indicators are counted) described in the paper under consideration. Likewise, a file record for a process also embraces the conditions of that process, for example temperature, pressure. This is a valuable and powerful synthetic device and makes possible economic mechanised searches of the kind "Polyamide as coating material", "Phosphoric acid as etching agent" with 100% precision, without overloading a thesaurus with corresponding terms for the unanalysed, composite concepts and without introducing "relations" of the kind:

*Polyamide;*
  Related terms:
    Polyamide as coating material
    Coating material

*Phosphoric acid;*
  Related terms:
    Phosphoric acid as etching agent
    Etching agent

On the other hand, the pure logics of concept connections (logical AND, inclusive OR, exclusive OR, OPTIONALLY, NOT) are exactly represented for each file record in the TOSAR graph on the syntactic tape. Also the simultaneity or sequence in time of processes, possibly under different conditions of temperature, pressure, reaction time etc. as well as the typical association of several substances in various mixtures etc. are exactly represented in this graph in a predictable form and thus made available for the mechanised search.

Due to the considerable expenditure involved in drawing and storing TOSAR graphs, however, one will limit their employment to fields where it is precisely the aforementioned concept connections that are of utmost importance. This is for example the case with the literature in the polymer field.

Which are the demands that would have to be made on thesaurus terms or on a classification if it is to achieve an equal degree of fidelity of the representation of topics of the kind demonstrated above? First of all, one would have, by such a term, to link with each other those concepts that belong to one and the same sequence of processes, because they are obviously more closely related with each other than with the objects of another sequence of processes described in the same document. A closer linkage would have to be established between those concepts that are involved in a certain process in such a way that they represent the initial and the final states of this process. Those concepts that also belong to the same sequence of processes but do not immediately succeed each other in time would have to be linked less closely. Those concepts that occur in the same point in time (as is the case, for example, with the various starting materials for a certain process) would have to be particularly closely linked by a suitably defined set of hypothetical terms. The components of a mixture are still more closely linked. Such a system of links would also have to express the logical compatibility of concepts. The sequence of concepts in time, particularly that of processes, would also have to be expressed.

It is hard to imagine that the representation of all these concept connections with a vocabulary of any kind, *be it a thesaurus or a classification*, can be rendered sufficiently logical, complete and transparent, so that one can in practice rely on the essentials of a document to be consistently represented, unless such an imagined system assumes a graphical shape. Otherwise the rules for such a hypothetical system would be too complicated, thus rendering the reliability of the searches questionable. Many failures of link usage reported in the literature apparently resulted from the lack of efficiency of the links investigated and from lack of clarity of the rules for the assignment of these links (cf. 17).

## 7. Conclusion

If the inherent features of both the thesaurus and the classification approach are compared it becomes obvious that they can effectively complement each other. In combining them it is possible to eliminate their specific drawbacks and to preserve their specific advantages. Such an approach has already been pursued by several authors (18, 19) and was reviewed by Lancaster (20). A particular variant of this conception was described in this paper. Even medium sized computer facilities suffice for the realization of such an integrated conception if one is ready impartially to familiarize oneself with the principles of both these approaches that are still widely considered opposite and incompatible.

The establishment and operational use of such an integrated system is doubtlessly more expensive at the input stage than is the case with a less sophisticated, weaker system. However, this extra expenditure serves the purpose of assuring a high *fidelity* of the representation of concepts and their relations in the store, as well as the *predictability* of their modes of expression. Thus, the resources are created to which one will increasingly have to resort under the constraints of the future, in which one will be confronted with a continuously growing store and a likewise increasing inquiry specificity, which in itself is a natural consequence of the increasing specialization in science and technology. The future will also lead to an increase in the inquiry frequency, which is a desirable consequence of the (hopefully!) increasing popularity of the system.

Exaggerated parsimony at the storage stage will inevitably impair the quality of the search responses and, thus, seriously reduce the survival power of the system. For, lacking *fidelity* will lead to a high and continuously increasing (and often repetitive) expenditure in weeding out masses of irrelevant documents, and, in an advanced stage, to the rejection or omission of an increasing number of such ballast-including inquiries.

Lacking *predictability* will result in a high and continuously increasing, even if often latent, proportion of loss of relevant information. In weaker systems, *expenditure at the search stage* will increase particularly rapidly if such systems are employed for a particularly large literature coverage (e. g. on the global scale) and over an extended period of time. If such a retrieval system has later to be abandoned either wholely or in part, a circumstance which may manifest itself, for example, in a restriction in the search frequency or in a limitation of the file to be searched (e. g. by resorting to SDI-services), then this is equivalent to a particularly heavy loss of stored, relevant and, hence, valuable information as well as of reputation.

Thus, it is not only from the viewpoint of retrieval system performance, as expressed in terms of recall and precision, that those systems that combine the thesaurus and classification approach and avail themselves of powerful synthetic devices, are the most promising, but they are also superior as regards the long-term economic aspect.

## References

The selection of references has been limited to most recent literature)

1 Fugmann, R.: *The Theoretical Foundation of the IDC-System: Six Postulates for Information Retrieval.* In: ASLIB Proc. 24 (1972), No. 2, p. 123–138.

2 Robertson, S. E.: *In Defence of Relevance.* In: J. ASIS 25 (1974), No. 3, p. 208.

3 Fugmann, R.: *On the Role of Subjectivitiy in Establishing, Using, Operating and Evaluating Information Retrieval Systems. Treatise II on Information Retrieval Theory.* In: Storage and Retrieval 9 (1973), p. 353–372.

4 Kemp, D. A.: *Relevance, Pertinence and Information System Development.* In: Inform. Storage and Retrieval 10 (1973), p. 37–47.

5 Yovits, M. C., Whittemore, J.: *A Generalized Conceptual Development of the Analysis and Flow of Information.* In: J. ASIS 24 (1973), No. 3, p. 221–231.

6 Mitroff, I. I., Williams, J., Rathswohl, E.: *Dialectic Inquiring Systems: A New Methodology for Information Science.* In: J. ASIS 23 (1972), No. 6, p. 365–378, esp. p. 372.

7 Wersig, G., Meyer-Uhlenried, K. H.: *Versuche zur Terminologie in der Dokumentation II.* Nachrichten für Dokumentation 20 (1969), p. 199.

8 Mills, J.: *Progress in Documentation.* In: J. Doc. 26 (1970), p. 123.

9 Ranganathan, S. R.: *Prolegomena to Library Classification.* London: ASIA Publishing House 1967.

10 Bhattacharyya, G.: *Chain Procedure and Structuring of a Subject.* In: Libr. Sci. 9 (1972), p. 585–636, esp. p. 609.

11 TDCK Circular Thesaurus System; The Hague, Netherlands Armed Forces Technical Documentation and Information Center 1963.

12 Rolling, L.: *The Role of Graphic Display of Concept Relations in Indexing and Retrieval Vocabularies.* Brussels: European Atomic Energy Community 1965.

13 Jansen, R.: *Sachverhaltsdokumentation und Thesaurusentwicklung.* Proc. Intern. Conf. on General Principles of Thesaurus Building. Warsaw 1970, p. 15–30.

14 Soergel, D.: *Klassifikationssysteme und Thesauri.* Frankfurt: Deutsche Gesellschaft für Dokumentation 1969. 224 p.

15 Fugmann, R., Nickelsen, H., Nickelsen, I., Winter, J. H.: *Representation of Concept Relations Using the TOSAR system of the IDC: Treatise III on Information Retrieval Theory.* In: J. ASIS 25 (1974), No. 5, p. 287–307.

16 Henrichs, N.: *Dokumentationsspezifische Kennzeichnung von Deskriptorbeziehungen.* Proc. Conf. German Documentation Society. Bad Godesberg 1974. In print.

17 Coates, E. J.: *Some Properties of Relationships in the Structure of Indexing Languages.* In: J. Doc. 29 (1973), p. 390–404.

18 Wall, R. A.: *Indexing Language Structure for Automated Retrieval.* In: Inform. Storage and Retrieval 9 (1973), p. 607–617.

19 Gilchrist, A.: *Classification in the Building Industry.* In: J. Doc. 28 (1972), p. 296–320, esp. p. 304.

20 Lancaster, F. W.: *Vocabulary Control for Information Retrieval.* Washington DC: Information Resources Press 1972, p. 43–69 (including Aitchison's Thesaurofacet, Brenners Subject Authority List of the American Petroleum Institute, G. London's thesaurus for the field of meteorology, Barhydt's and Schmidt's Information Retrieval Thesaurus of Education Terms and the Exploration and Production Thesaurus).