Thesaurus Analysis for Updating*

ROBERT F. SCHIRMER
Information Systems Division, Secretary's Department, E. I. du Pont de Nemours & Company, Inc., Wilmington, Delaware

Received November 29, 1966

A thesaurus requires updating as new concepts are introduced. Techniques have been developed for updating and providing the necessary vocabulary control as the Thesaurus content changes. This paper discusses the results and significance of a study of nonchemical terms at Du Pont as being of possible interest to other information systems based on similar technologies.

In 1964, nine information centers in Du Pont began consolidation of their activities in the Information Systems Division of the Secretary's Department. Together these indexes had 50,000 reports in storage and processed over 3000 inquiries per year. It was found desirable to merge these nine indexes together with their thesauri into one consolidated system in order to establish a more efficient operation at a lower cost (*1-8*).

Other organizations have successfully created thesauri either through the selection of terms believed to be relevant or through the consolidation of terms previously in existence (*9-15*). In most of these cases the thesauri were based on theoretical rather than analytical conclusions. For example, in May 1960, the Defense Documentation Center, then known as ASTIA, issued a thesaurus in which 70,000 subject headings were reduced to 7000 terms. The Engineers Joint Council published, in June 1964, a thesaurus of 10,500 terms which were selected from 119,000 terms. Most organizations have not had the opportunity to study their question patterns or density of postings extensively as a basis for revising or creating their thesauri.

Du Pont was fortunate in being able to analyze its past operating experience in order to build a consolidated Du Pont thesaurus. Du Pont had up to 15 years of experience in one system, had a record of thousands of questions asked of the different systems, and had files which were in machine-processible form. This provided Du Pont the opportunity to analyze its needs and develop the best consolidated thesaurus in addition to establishing guidelines for future thesaurus updating, vocabulary control, and indexing procedures.

The study was limited to those concepts which could be classified as nonchemical. For purposes of the analysis, chemical concepts were defined as those concepts representing chemical compounds or chemical mixtures. All other concepts, such as processes, properties, devices, adjectives, etc., were considered as nonchemical. For the storage and retrieval of chemical compounds, it was decided to use the Chemical Abstracts Services' Chemical Structure Storage and Search System. CAS and Du Pont have been engaged in a joint research project in the development of a topological system for indexing and storing chemical structures. This system provided the mechanism for consolidating all like compounds into one representation regardless of the nomenclature or fragmentation system which was previously used by any of the nine departmental systems.

Included in Du Pont's study of nonchemical terms were an analysis of the degree of overlap of identical concepts between the nine information systems, the type and frequency of terms used in searching, the relationship between terms used in indexing *vs.* those used in searching, and the density of postings of both indexing and searching terms.

DISCUSSION

Number and Types of Terms. The initial phase of the study was to determine the number and type of different nonchemical concepts existing in the nine information systems. The first step was to spill off all the existing terms with their English descriptions from the computer files and sort them into either chemical or nonchemical categories. This provided a total of 25,700 nonchemical terms. The nonchemical terms were submitted to a digit-by-digit computer comparison of the English descriptions. Terms from two of the indexes were excluded from the computer comparison as they had, at one time, operated jointly with other indexes. Having shared a common vocabulary, it was felt that a high degree of vocabulary overlap would exist between these systems and would, therefore, bias the outcome of this particular analysis.

The analysis of the terms from the other seven systems showed 7500 terms to be exact duplicates representing 2790 different nonchemical concepts, as shown in Table I. In two indexes, 1650 of the concepts were duplicated, while 20 concepts were duplicated in all seven indexes.

A further manual analysis of all the nonduplicated terms for word formats such as singular, plural, word ending, hyphenization, etc. showed that an additional 11% of the nonchemical concepts could be combined into the 2790

* Presented before the Division of Chemical Literature, 151st National Meeting of the American Chemical Society, Pittsburgh, Pa., March 1966.

Table I. Frequency of Duplicated Terms

Number of Duplications	Number of Terms
2	1650
3	660
4	280
5	120
6	60
7	20
Total	2790

duplicated concepts. No effort was made at this time to combine synonymous terms such as "changing" and "modifying," "blending" and "mixing," or "cutting" and "slitting."

In the other two departmental systems omitted from the computer analysis, it was determined that 2850 of their 3300 nonchemical concepts were duplicated in one or more of the other vocabularies.

Subtracting the computer-analyzed duplicates, the manually analyzed duplicates, and the two vocabulary overlaps, the total number of different nonchemical terms was determined to be 15,310 (Table II).

Table II. Number of Different Nonchemical Concepts

Total Nonchemical Concepts (9 Systems)		25,700
Less:		
Computer-Analyzed Duplicates (7 Systems) (7500 – 2790)	4,710	
Manually Analyzed Duplicates (9 Systems) (11 × 25,700)	2,830	
Vocabulary Overlaps (2 Systems)	2,850	10,390
Total Different Nonchemical Concepts		15,310

Each of the duplicated terms was reviewed and classified into nine categories (Table III): proper names, plants and geographical locations; trade and code names; materials, mixtures, and states of materials; energy and forms of energy; devices, equipment, tools, and components; shapes and forms; processes, operations, and technologies; qualities, properties, and conditions; and adjectives and modifiers. Each of the categories was tabulated with the results shown in Table IV. The largest categories in descending order were devices, processes, materials, and qualities. These four categories accounted for 83% of all the concepts.

It was concluded from Phase I that there was a high degree of overlap of nonchemical concepts common to two or more systems and that for indexing purposes, the majority of these concepts could be classified as devices, processes, materials, or qualities.

The study to this point included only those terms used for indexing. Unanswered were the questions: "Were the same conclusions valid for terms used in searching?" "Was there an overlap in search terms common to two or more indexes?" "Was there a pattern as to the categories into which search terms could be classified?"

Terms Used in Searching. In the second phase of the study, an analysis was made of nonchemical concepts used in 2100 computer searches of the three largest systems over a 1-year period. Although many manual searches were also made of these indexes, only computer searches were selected for the study, as the data were more accessible. These computer searches were the more complex searches, as past experience indicated that about 50% of all searches were made manually.

Table III. Definitions of Term Categories

Category 1, *Proper Names*: Plants, Geographical Locations.
Examples: American Can Co., Beaumont Works, Texas, Germany

Category 2, Trade or Code Names.
Examples: "Carbowax,"® "Lucite,"® "Orlon"®

Category 3, *Materials:* Mixtures, States of Materials.
Examples: Chemicals, Ceramics, Costs

Category 4, *Energy*: Forms of Energy.
Examples: Heat, Electricity, Ultraviolet Radiation

Category 5, *Devices*: Equipment, Tools, Components.
Examples: Dryers, Drills, Mandrels

Category 6, *Shapes*: Forms.
Examples: Bars, Interfaces, Chips

Category 7, *Processes*: Operations, Technologies.
Examples: Spinning, Polymerization, Mathematics

Category 8, *Qualities*: Properties, Conditions.
Examples: Elasticity, Temperature, Velocity

Category 9, *Adjectives*: Modifiers.
Examples: Blue, Cold, Dimensional

Table IV. Duplicated Term List of Nonchemical Term Categories

Categories	Number of Different Terms	%
Devices	750	27
Processes	680	25
Materials	530	19
Qualities	340	12
Adjectives	190	7
Proper Names	90	3
Shapes	90	3
Trade Names	90	3
Energy	30	1
Total	2790	100

The different nonchemical concepts searched from the three systems were tabulated and compared with the total number of nonchemical concepts in existence in each vocabulary. The results indicated that only 2010 of the 11,300, or 18%, of the nonchemical terms were used in searching (Table V). All of the 2010 nonchemical search terms were classified into the nine previously discussed categories (Table III), together with their frequency of use.

This analysis (Table VI) showed that proper names, trade names, shapes, energy, and adjectives accounted for only 10% of the different nonchemical concepts searched and were used only 8% of the time. Devices, qualities, materials, and processes represented 90% of the different concepts searched and were used 92% of the time.

In total, the results showed that the 2010 nonchemical concepts were used more than 5300 times. Fifty-two per

Table V. Nonchemical Concepts Searched

System	Total Nonchemical Concepts in Vocabulary	Different Nonchemical Concepts Searched	% of Nonchemical Vocabulary Searched
I	3,700	760	20
II	4,000	545	14
III	3,600	705	20
Totals	11,300	2010	18

Table VI. Frequency of Search Terms by Category

Frequency	Proper Names	Trade Names	Shapes	Energy	Adjectives	Devices	Qualities	Materials	Processes	Total	%	
1	1	32	18	13	41	216	182	192	341	1,036	52	
2	0	1	8	6	24	77	67	74	148	405	20	
3	0	0	7	6	8	36	48	42	56	203	10	90%
4	0	0	1	3	4	10	21	21	33	93	5	
5	0	0	1	2	2	7	13	18	27	70	3	
6–10	0	2	1	0	5	13	29	32	46	128	6	
11–49	0	0	0	2	1	7	13	26	26	75	4	
Total	1	35	36	32	85	366	373	405	677	2,010		
%	0	2	2	2	4	18	19	20	33		100	
			10%					90%				
Frequency of Terms Searched	1	47	71	92	193	745	1,014	1,347	1,816	5,326		
%	0	1	1	2	4	14	19	25	34		100	
			8%					92%				

cent of the terms occurred only once, 72% occurred once or twice, and 90% were used five or less times.

These same data, when expressed in terms of the ratio of nonchemical terms searched per question by the nine categories, are shown in Table VII. Proper names were used only once in the 2100 questions, while trade names were used once in every 45 questions. On the other end of the scale, devices were used once in every three questions; qualities, once in every two questions; materials, once in every 1.5 questions; and processes, once in every one question.

Search terms were also studied with respect to the 2790 duplicated indexing terms which were discussed in Phase I. The search terms were analyzed first to determine if they were exact matches for concepts on the duplicated term list. If not, they were analyzed to determine if they could be easily modified to coincide with concepts on the duplicated term list. For example, the concept "fiber blend" could be split into the two concepts "fibers" and "blends." Likewise "wear tests" could be changed into "wear" and "tests," "wire coatings" could be divided into "wires" and "coatings," and "fabric uniformity" could be split into "fabrics" and "uniformity."

The results of this analysis of search terms from the three largest systems revealed that 89% of the concepts from System I, 96% of the concepts from System II, and 80% of the concepts from System III or an average of 89% of all nonchemical concepts used in searching were contained in, or could be easily converted to, the 2790 concepts on the duplicated term list (Table VIII).

Table VII. Ratio of Terms Searched Per Question by Category

Category	Ratio
Proper Names	1:2100
Trade Names	1:45
Shapes	1:30
Energy	1:23
Adjectives	1:11
Devices	1:3
Qualities	1:2
Materials	1:1.5
Processes	1:1

Phase II disclosed that only a small percentage of nonchemical concepts was used in searching and that the majority of these concepts also fell into the categories of devices, qualities, materials, and processes. These results, coupled with the 89% conformity of search terms to concepts on the duplicated term list, indicated that a relatively small, well-defined thesaurus would adequately meet the requirements of the consolidated indexes.

Results for systems with vastly different parameters could not be concluded, because no data on such systems were available within Du Pont.

Density of Postings. The third phase of the study analyzed the density of direct postings in the term/document files of the respective systems. Because this statistical analysis was determined from the existing computer files, no attempt was made to distinguish between chemical and nonchemical concepts. As shown in Table IX, the study disclosed 47% of all terms to have only one direct posting, 22% to have more than five postings, 13% to have more than ten postings, 8% to have more than twenty postings, and only 4% to have more than fifty postings.

Table VIII. Comparison of Search Terms *vs.* Duplicated Terms

	%
SYSTEM I (760 Terms)	
Exact Duplicates	82
Converted Duplicates	7
Total Duplicates	89
SYSTEMS II (545 Terms)	
Exact Duplicates	88
Converted Duplicates	8
Total Duplicates	96
SYSTEM III (705 Terms)	
Exact Duplicates	51
Converted Duplicates	29
Total Duplicates	80
TOTAL OF THREE SYSTEMS (2010 Terms)	
Exact Duplicates	76
Converted Duplicates	13
Total Duplicates	89

Table IX. Density of Postings

POSTINGS	INDEX TERMS (Incl. Chemicals)	SEARCH TERMS (Excl. Chemicals)
1	47%	2%
More than 1	53%	98%
More than 5	22%	92%
More than 10	13%	87%
More than 20	8%	75%
More than 50	4%	51%

In analyzing the density of direct postings to search concepts, 51% of the nonchemical concepts were found to have more than 50 direct postings, 75% more than 20 postings, 87% more than 10 postings, and 92% more than five postings.

Since the analysis was restricted to those concepts used in computer searches, it is expected that these percentages would have been somewhat lower if concepts used in manual searches had been included. Searches which are suitable for manual manipulation usually contain one or more highly selective terms—*i.e.*, relatively few postings.

Although the analysis of indexing terms included postings to chemicals, it is believed that there would not have been any significant difference in the results if the chemical terms had been omitted.

Guides for Indexing. Based on the analysis of the types of concepts used in searching, it was concluded that proper names, energy, shapes, and adjectives were seldom used and therefore, should not normally be indexed. When necessary to use adjectives, they should be bound.

Since trade names were rarely searched, "see" references should be established to their component materials wherever possible. Concepts should be nonsynonomous with unnecessary overlaps and shadings of meanings avoided. Nonchemical concepts should be indexed as nouns or gerunds when possible.

From experience in searching, for example, it was found that when indexing the information that "corrugated boards were coated with a blend of 'Elvax' 260 using an 'Egan' curtain coater," indexing terms selected should be "coating," the gerund; "corrugated boards," with the bound adjective; "blends," the plural noun; and "curtain coaters," the bound concept. "Elvax" 260 should be indexed by its chemical composition, "vinyl acetate ethylene copolymer," and the proper name, "Egan," should not be indexed.

SUMMARY

Based on operating experience and studies of the data, it was found that a high degree of overlap existed of nonchemical concepts common to two or more systems. Of these concepts, 83% were categorized as devices, processes, materials, and qualities. Results indicated that only 18% of the different nonchemical concepts were used in searching and that 89% of the search concepts conformed with the list of duplicated terms common to two or more systems. Ninety per cent of the search concepts fell into the categories of devices, processes, materials, and qualities. It was also found that 51% of the computer search terms had more than 50 postings, while 92% had more than five postings.

These results disclosed that only a small percentage of the indexing terms was actually used in searching and that the vast majority of these concepts could be confined to a relatively small vocabulary of nonchemical concepts. The high density of postings to computer search terms indicated that the concepts used in searching were the more frequently used indexing terms. The analysis also showed that most of the search concepts could be classified as devices, processes, materials, and qualities.

The results of this study have furnished the basic data for establishing guidelines for creating a consolidated thesaurus of the nine departmental systems. It was found that for Du Pont's operation certain concepts such as devices, processes, qualities, and materials were more important than others and that relatively few different nonchemical concepts were used in searching. These facts were concluded based on studying such diverse systems as Textile Fibers, Explosives, Engineering, and Marketing Research. Indexers have been instructed to index proper names, energy, shapes, and adjectives with discretion; to bind adjectives whenever possible; to establish "see" references for trade names when applicable; and to avoid overlaps and shadings of meanings of concepts.

A small consolidated thesaurus of mutually exclusive concepts will provide the necessary mechanism for communications between systems in order that document references from all nine indexes can be retrieved in answer to a single search.

It is estimated that 18 man-months will be required to edit the 25,700 nonchemical terms from the nine vocabularies into a consolidated thesaurus. This effort is considered a worth-while investment, as a small vocabulary is less complicated, easier to use and maintain, and less costly for machine processing. The consolidated thesaurus and consolidated term/document or inverted files currently under preparation will provide the centralized indexes with an efficient, low cost, one inquiry–one search method of operation.

ACKNOWLEDGMENT

The author wishes to gratefully acknowledge the assistance of Thomas E. Boyle, Jr., of E. I. du Pont de Nemours and Company, Inc., in reviewing and classifying the nonchemical terms.

LITERATURE CITED

(1) Word Correlation and Automatic Indexing, Council on Library Resources, April, 1960.
(2) Herner, S., Johanningsmeier, W. F., "Information Storage and Retrieval: Is it Working?," *Chem. Eng. Progr.* **61**, 23 (1965).
(3) Holm, B.E., Rasmussen, L. E., "Development of a Technical Thesaurus," *Am. Document* **9**, 184 (1961).
(4) Holm, B. E., "Techniques and Trends in Effective Utilization of Engineering Information," *ASLIB Proc.* **17**, 134 (1965).
(5) Speight, F. Y., "Procedures for Revision of the Thesaurus of Engineering Terms," Engineering Joint Council Memorandum to W. M. Carlson, Draft September 23, 1964.

(6) Wall, E., "Final Report—First Revision of the Thesaurus of ASTIA Descriptors," AD 278168, Armed Services Technical Information Agency, 1962.
(7) Wall, E., "Study of Engineering Terminology and Relationships Among Engineering Terms," AD 432231, Engineers Joint Council, New York, N. Y., August 1, 1963.
(8) Wall, E., "Information Retrieval Thesauri," Engineers Joint Council, New York, N. Y., November 1962.
(9) "Chemical Engineering Thesaurus, "American Institute of Chemical Engineers, 1961.
(10) "DDC Authorized Descriptors," Defense Documentation Center, July 1964.
(11) "Euratom, Thesaurus," European Atomic Energy Community, Euratom, 1964.
(12) "Thesaurus of Descriptive Terms and Code Book," 1st. ed., Department of the Navy, Bureau of Ships, December 1963.
(13) "Thesaurus of Engineering Terms," 1st. ed., Engineers Joint Council, May 1964.
(14) "Thesaurus of Descriptors," tentative edition, U. S. Department of the Interior, Bureau of reclamation, October 1963.
(15) "Thesaurus of ASTIA Descriptors," Armed Services Technical Information Agency, 1960.

Biomedical Information Retrieval: A Computer-Based System for Individual Use

C. N. GILLIS*
Department of Pharmacology, Yale University School of Medicine, New Haven, Connecticut

Received December 8, 1966

A computer-based system has been developed for the retrieval of information from individual collections of abstracts, reprints, and other information relevant to individual research interests. The system and its operation are described.

The problems involved in storage and retrieval of biomedical information are receiving increasing attention from specialists in the field. There exist currently several centralized, computer-based systems that offer, on a service basis, access to the current literature. To use such services—for example, MEDLARS: Medical Literature Analysis and Retrieval System—requests for searches of the stored information must provide key words or phrases relevant to the subject; the occurrence of these in the title of a paper results in the title and journal reference being printed as output. Although difficulties exist in adequately indexing papers for such systems (*1*), they can be invaluable in searching the literature for information on a particular topic.

Much more frequently, however, investigators refer to their own collection of index cards, reprints, and notes, based on papers that are relevant to their particular interests. Over a period of several years, ready and reliable access to specific information in such collections becomes more difficult to achieve. Many people attempt some form of filing of index cards based on broad subject categories. Another widely used system employs cards whose edges are punched at locations corresponding to numbers assigned to topics mentioned in the paper. This system allows reasonably adequate cross referencing of articles, but again a problem arises in searching adequately through an ever-increasing number of such cards. Over the past few years, the author has used a system utilizing the edge-punched cards referred to above. This paper describes a recent adaptation of the system that allows computer-aided searching of the stored information.

Papers read are assigned an appropriate selection of up to 10 primary, 10 secondary, and six tertiary two-digit numbers, each corresponding to one of the topics listed in Table I. The primary topics encompass broad areas of interest, while the secondary and tertiary topics are those dealing with (at least for the author) increasingly specific subject material. The list of topics (Table I) and their corresponding numbers, although quite large is, in practice, easily committed to memory with continued use. As papers are read, notes may be made on index cards which are numbered consecutively and filed in numerical order. The complete information on each paper (termed the reference set) consists of the numbers categorizing the paper, its title, authors, and journal reference. Each reference set is typed on a coding sheet together with the same number as the corresponding index card with notes made at the time the paper was read. If no card was prepared when the paper was read, reference sets

* Work done during the tenure of an Established Investigatorship of the American Heart Association. The cost for development of the computer program was borne by a grant from the Natural Science Foundation (GP-4774) to Yale University. Other expenses were met by Grants H-7249 from the National Heart Institute and O-65-46 from the Life Insurance Medical Research Fund.