Deep Indexing Technical Reports*

By PAUL LOGUE
Information Center, Research & Engineering Division,
Monsanto Chemical Company, St. Louis, Missouri
Received May 9, 1962

INTRODUCTION

Monsanto's purpose in devising a "deep index" for technical reports was to increase the access to and the utilization of its internally generated technical reports, which represent an average investment of $20,000 each. It was felt that although the primary objective of the report already may have exerted its effect on the conduct of the business, auxiliary benefits can accrue if the contents of these reports are kept constantly before and readily accessible to the men working at the "leading edge" of scientific progress for the company. We began with the current situation and worked backward through the previously issued reports. The task is now about half completed.

Although individual engineers had previously employed coordinate indexing for small personal files, this approach had not been used by Monsanto at a department or division level until it was initiated in 1958 for indexing chemical economics information in the Development Department of the Research and Engineering Division. Interested members of the department read about 25 journals in which they circled locators in red for later use as "handles" to retrieve the information. Auxiliary locators could be written in the margin if implied or if customarily used by the reader in searching for information. The marked journals were clipped by the reader's secretary, who also typed a "data input slip" listing the circled terms, the magazine reference, and the "reader source." This "input slip" (Fig. 1) was reviewed by the "editor-librarian," who assigned a sequential "accession number" both on the "input slip" and on the original clipping. The clippings were filed by accession number in an upright file drawer and the "input slip" became the "tracing sheet" for the key-puncher.

UNITERM NO. P 3996
DATE OCT 27 1959
FILE: CHEMICAL ECONOMICS LIBRARY, R. & E. Division - Development Department
UNITERM DATA PROCESSING REQUEST

1	TITLE: FIRESTONE TO MAKE NYLON	DOCUMENT LOC/PHONE
	SOURCE: CHEM ENG NEWS — AUTHOR — DATE: 9/28/59	
2		
3	UNITERMS: FIBERS TIRE CORD VIRGINIA CONSTRUCTION NYLON POLYPROPYLENE CORAL POLYISOPRENE RUBBER Production	
	PRODUCT	
4	COMPANY: FIRESTONE TIRE & RUBBER	
5	IND. SIC NO	
6	SUBMITTED BY: MOWRY — TEL. NO. (G.O. Only): 2397 — ADDRESS (Other than G.O.)	

NOTE: DO NOT COMPLETE SHADED AREAS

Fig. 1—Typical data input slip.

The words themselves, not coded, were used as locators. These were inverted on an IBM 650 computer to yield a printout of an alphabetical index (Fig. 2) of terms with pertinent accession numbers arrayed under each term.

FOAM

860	81	632	1533	1244	745	2596	1387	828	519
1320	831	832	2063	2584	1805	3966	4487	1868	659
1620	1771	1182	2523	3964	3055	4176	4677	2048	859
1830	2591	1962	3983	4874	4105	4956	5047	3248	969
2480	4851	2932	4523	5114	4525	5736	6067	6368	1529
3520	5711	5162	6173	5344	7515	6866	6647	6588	1599
4170	7991	5932	6563	6144	7865	8406	8747	7678	2029
5150		7182	6943	6424	8065	9326		7848	2709
5500		9102	7583	7234	9415	9336		8378	5709
5710		9522	7853	7524		9726		8838	7069
6370			8023	8144				9658	8679
7850			8123	8284				9728	8829
8820			8503	8874					
9660			8663	9144					
			8783						

Fig. 2.—Typical inverted index machine-printed on 5 × 8 cards.

Searching was accomplished by manual comparison of the accession numbers under selected terms. Searches used 5 × 8 term cards in the library index or dual dictionaries (Fig. 3) at remote points very much like those described by Costello.[1]

The chemical economics file accumulated 11,500 documents in two and a half years. The input level reached 100 documents per week and required a staff of two persons to convert the terms circled by interested readers into punched cards for processing on an IBM 650 computer which inverted the file to yield a printed dual dictionary index of locators for manual searching. The same personnel not only edited the input terms but also provided retrieval of documents from the numerically sequenced file on request. Although time spent on the input side was disproportionately high at first, it finally smoothed out to about 50¢ for input and 50¢ for handling inquiries and retrieval. On this basis the input cost per document filed was $1.95 exclusive of the time of the reader who circled the locators, or terms, initially or the time of his secretary who typed the input slip.

INDEXING TECHNIQUES

The successful execution of this relatively simple indexing system provided background and experience for

*Presented before the Division of Chemical Literature, ACS National Meeting, Washington, D. C., March 23, 1962.

CHEMICAL ECONOMICS LIBRARY

COORDINATE INDEX

DEVELOPMENT
DEPARTMENT

RESEARCH AND ENGINEERING DIVISION

800 North Lindbergh Boulevard Saint Louis 66, Missouri

Monsanto

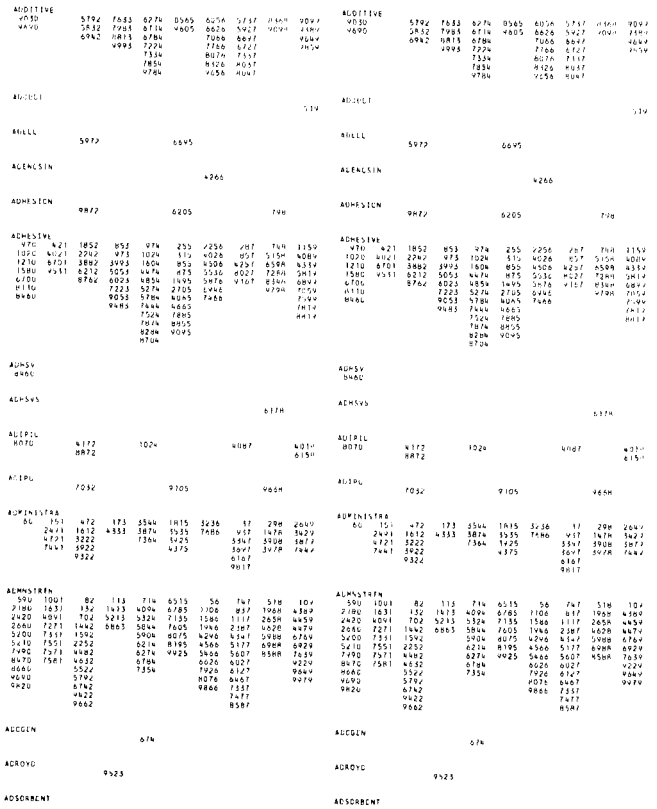

Fig. 3.—Dual dictionary of coordinate concept index terms.

the more complicated task of generating a deep index for technical reports.

To provide additional guidance in the use of links (topical subdivisions within a document of related subject matter) and roles (syntactical indicators of the significance of terms in context)[2, 3, 4, 5] the services of a consultant were engaged. It was decided that current reports would be indexed in depth directly from the report, but that all past reports would be indexed using only the information furnished in the report announcement, index and abstract. Three college juniors who had completed 3 or more years of chemistry were trained by the consultant in the technique of coordinate indexing and were assigned the task of indexing the backlog of 22,700 reports using a form as illustrated in Fig. 4. They actually indexed 8,500 reports from the report announcements during the summer at a direct cost of $3,328, or $0.39 per report. Indexing was carried to an average depth of 12.2 locators per report, making the indexing cost $0.032 per locator. The actual time required for the indexing was 1440 hours. The average indexing rate was 5.9 reports per hour, or 0.169 hour per report.

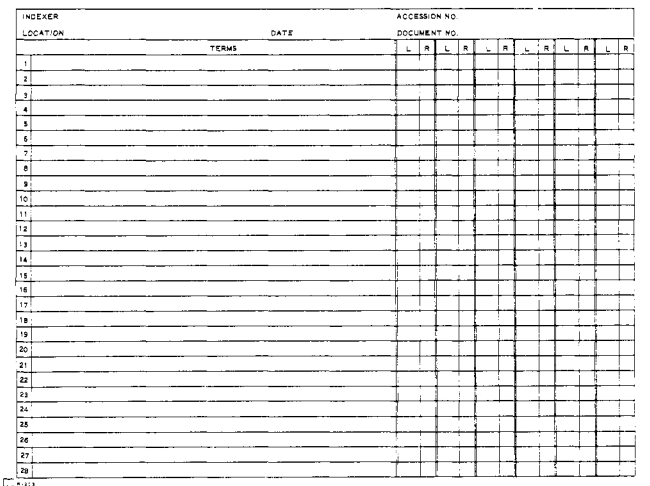

Fig. 4.—Tracing sheet used for indexing technical reports.

After some experience a form was developed which can accommodate both internal and external technical information (Fig. 5). On this form emphasis is placed on

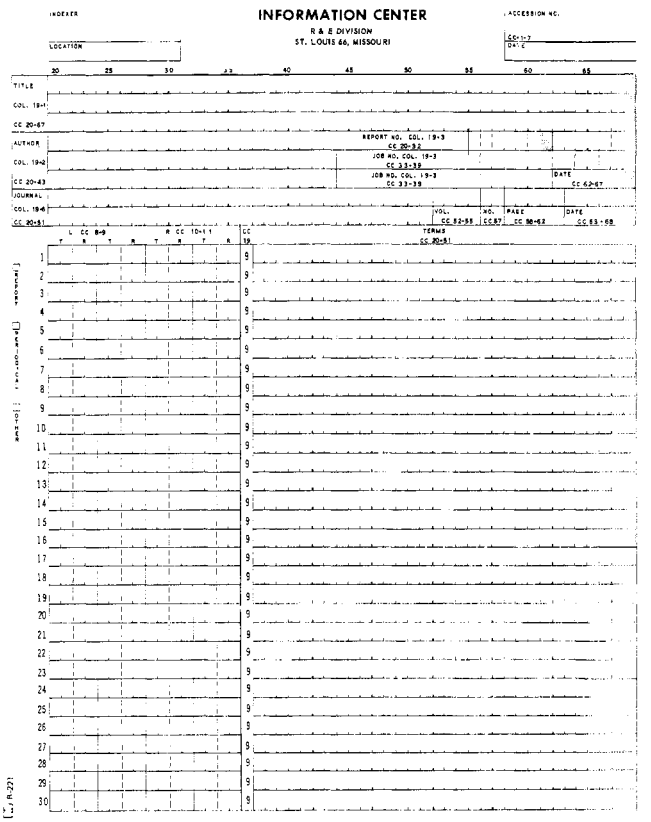

Fig. 5.—Revised tracing sheet for indexing technical reports, translations, manuscripts and external literature.

ordering the information for ease of transfer to assigned columns on punched cards.

The roles chosen by Monsanto were assigned a two-digit code number in an ascending sequence so that, as arranged by the machine, they would be printed in a logical order and interpreted in a stylized form as illustrated in Fig. 6.

ROLE	INTERPRETATION	ROLE	INTERPRETATION
08	STUDY OF	33	REACTANT
10	DESIGN OF	45	ENVIRONMENT
12	PATENT ON	54	SPECIAL AGENT
13	REVIEW OF	56	BY PRODUCT
14	GOVERNMENT	62	IND. VARIABLE
15	BUSINESS INFORM.	69	DEP. VARIABLE
21	OF. IN. ON	71	USING. USED FOR
27	PRODUCT. OUTPUT	90	(leave blank)

Fig. 6.—Stylized interpretation of the role codes used in the bibliographic abstract.

INSTRUCTIONS FOR USE OF ROLES IN INDEXING MONSANTO REPORTS

A. ENVIRONMENTAL OR SOURCE ROLES

08. *Study of. Investigation of*

a. Must be a real contribution, not a review. Role 08 deals with an operation, process, property, equipment, experiment design, instrument, use-class, etc. It is never used on substances. Role 08 is employed on translations of foreign articles and manuscripts submitted for publication, both of which are treated the same as Monsanto reports.

b. To differentiate between the levels of study or investigation, six adjectives will be used: theoretical, mathematical, laboratory, pilot-plant, semiworks, and plant-scale. The terms *theoretical* and *mathematical* describe prelaboratory studies and investigations.

c. A number of laboratory reports will contain suggested procedures and tentative processes. If a laboratory report contains either, index appropriate topics by either the term *suggested procedure* or *tentative process* and treat them as adjectives.

d. In describing the synthesis of organic compounds or preparation of inorganic compounds, use the term-role *08 synthesis* or *08 preparation*. If the name of the process can be specifically identified, use it in addition in both Roles 08 and 71, as:

08 synthesis	or	08 preparation
08 amination		08 calcination
27 propylamine		27 sodium tri-polyphosphate
71 amination		71 calcination

e. The term *preparation* may also be used to indicate the physical mixing or blending of two materials. If *preparation* is used in the chemical sense, then the compounds in question are indexed in Roles 33 and 27; if those being preparaed are being mixed, then use Role 21 for those materials.

f. Where the chemical reaction is a polymerization or copolymerization, do not use the term *synthesis*. Use *polymerization* if there is only one monomer being polymerized to form a homopolymer. Use the term *copolymerization* if there are two comonomers producing a copolymer. Similarly, for termonomers use the term *terpolymerization* and identify the products by three terms, with *terpolymer* as the last word in each of the three terms.

g. Although there may admittedly be some chemical reactions in such situations as emulsions, dispersions, colloidal suspensions, alloys, and amalgams, consider them as physical mixtures so that "the alloying of copper and nickel" will be indexed as:

08 alloying
21 copper
21 nickel

The analysis of a copper-nickel alloy is indexed as:

08 analysis
21 copper alloys
21 nickel alloys

h. In links which start with the term-role *08 analysis*, the material being analyzed for is in Role 21; all other materials present are in Roles 56 (since they are not the objects of the analysis) except solvents, which are in Role 45. This applies even if, from another point of view, what we are analyzing for might be considered an impurity and indexed from that other point of view with Role 56.

i. When the terms *performance*, *yield*, and *conversion* are used in connection with a chemical process, remember that *performance* refers to the equipment (*08 performance* with *21 converters*). *Yield* refers to the quantity of product obtained (*08 yield* with *21 polyethylene*). *Conversion* applies to amount of reactant reacted (*08 conversion* with *21 ethylene*).

j. Application studies or end-use investigations such as "Study of Applications for Polystyrene in Tooth Brushes in 1959" would be indexed as:

08 applications
21 polystyrene
71 tooth brushes
90 year 1959

10. *Design of. Design Data on. Design Information on. Drawing of Construction Information on. Plant Operation Information on*

Use Role 10 only on processes, operations and equipment, never on chemicals, materials or properties. Consider: "To what aspect of or topic concerning the substance should this role be assigned instead of using it on the substances?"

12. *Patent on (Meaning Claim of or Disclosure of in a Patent or Patent Application).*

13. *Book Concerning. Compilation Concerning. Manual Concerning. Instruction Book on. Literature Survey of. Review of*

14. *Governmental Relationships Concerning*

Use Role 14 only on processes, operations and equipment, never on chemicals, materials or properties. Consider: "To what aspect of or topic concerning the substance should this role be assigned instead of using it on the substances?"

15. *Business Information Concerning. Commercial Information, including commercial intelligence, legal matters, financial information, management matters, and personnel matters*

Use Role 15 only on processes, operations and equipment, never on chemicals, materials or properties. Consider: "To what aspect of or topic concerning the substance should this role be assigned instead of using it on the substance?"

B. TECHNOLOGY ROLES

27. *Output*

a. Product or coproduct from reaction.

b. When two things are conbined for the purpose of obtaining a desired end-product but it is not obtained, index the reactants in role 33 and the desired end-

product in role 27. When something other than the desired product is obtained, index both the desired product and what was obtained in role 27.

c. When reactions are performed but result in no identifiable end-product, index the reactant in role 33 but do not show anything in role 27.

d. In indexing polymer blends, such as "polyethylene-polypropylene blends," use two separate terms: *polyethylene blends* and *polypropylene blends*.

e. In indexing copolymers, such as "styrene-butadiene copolymers," use two terms: *styrene copolymers* and *butadiene copolymers*.

f. Complexes are treated the same as blends or copolymers: *cupric chloride complexes* and *ammonium sulfate complexes* as terms for "cupric chloride-ammonium sulfate complex."

g. In indexing metal compounds, use the words cuprous and cupric, ferrous and ferric, and nickelous and nickelic instead of copper, iron, and nickel. This will make it possible later on to merge *cuprous chloride* and *cupric chloride* to form one term *copper chlorides* if we wish. It would be impossible to do the reverse.

h. When the term *acids* is used, distinguish whether inorganic or organic acids are intended so that the terms *acids (inorganic)* and *acids (carboxylic)* may be used. Similarly for *acetates (salts* and *acetates (esters)*, etc.

33. *Reactant, Input*

a. Input to chemical reaction, reactant in chemical reaction.

45. *Reaction Medium, Atmosphere, Solvent, Vehicle, Dispersion Means, Environment, Support (as for a Catalyst)*

54. *Special Agent in a Reaction or Operation (Including Catalysts, Vulcanizing Agents, Color Stabilizers, etc.)*

a. A definition of "special." A special agent is a substance, either a compound, a mixture of compounds, or a product identified by trademark, which is added to a reaction or to a material to accomplish a change in the process or the characteristics or properties of the material. The special agent itself is indexed in Role 54. The use or particular type of special function which the substance is to perform is indicated by a term and the appropriate role is appended to the use term. As an example, in a 62-69 link, "the effect of cuprous chloride as a catalyst" is indexed:

62 catalysts
54 cuprous chloride

Another example: "Research on cupric chloride to be used as a catalyst in the polymerization . . ."

08 catalysts
54 cupric chloride
71 polymerization

56. *By Product, Impurity or Contaminant, Undesired Reaction Product or Waste, Product of a Side Reaction*

62. *Independent Variable Studied for its Effect on*

69. *Dependent Variable Studied for how it is Affected (62 and 69 must be used simultaneously)*

a. Normally 62 and 69 express the relationship of an independent variable on a dependent variable or the cause and its effect. In situations where two variables mutally affect each other, index both variables in both Role 62 and also in Role 69. This will express that, either alternately or simultaneously, they affect each other.

C. PSEUDOGRAMMATICAL ROLES

21. 2. Receiving a physical modification (not due to chemical change)

b. Passively receiving an operation (such as storage)

c. Object of the prepositions of, in, and on (possessive and locative)

d. If we are separating two materials, they are both in Role 21. If we are purifying or removing, the undesired material is indexed in Role 56 and that which is being purified in Role 21.

71. a. Using, by means of, by

b. Used to, used for, use later on for, applied to, applied for, for later use in

90. a. Adjectives

b. Proper names including names of companies, persons, competitors, divisions of the company, plant locations, and government agencies. (This includes all proper names except those used with environmental or source roles. Thus, information concerning a trademarked product would be denoted as, say, *21 Sterox* and a book on the AEC would be *13 Atomic Energy Commission.*)

A seven-digit term code to handle the involved engineering concepts and the complicated chemical terms was created by manually arranging in alphabetical sequence (according to *Chemical Abstracts* rules) the punched cards for: the terms from the chemical economics file mentioned above; the terms used by the Chemstrand Corporation (a Monsanto subsidiary); the "lead terms" of subject headings from the standard library cards used previously to index technical reports; and the locators from the A.I.Ch.E. Thesaurus. To this comprehensive word list of 27,804 terms the IBM 1401 computer assigned equal interval codes in a total range from 0 to 9,500,000. The last 500,000 numbers (from a total of 10,000,000) were reserved for an empirical chemical formula index. The generation of this term cost $3,740, or $0.134 per word.

This word list is now being refined into an acceptable index vocabulary and thesaurus. *Chemical Abstracts* terminology has been selected as "standard" and is departed from only in rare cases warranted by long established internal usage. Instructions are being developed for posting document numbers by machine from specific terms to generic terms based on hydrocarbon structure and reactive groups, to related terms, and to empirical formula.

The term cards were sorted alphabetically and in this grouped position were assigned the appropriate term code, which was gang-punched into each group of cards. At the same time, a dictionary card containing only the term code and its alpha-numeric term equivalent is being generated which will be used as a master card for generating an alphabetically ordered vocabulary of terms used. From this store of punched cards the indexer chooses the appropriate cards and passes them with the tracing sheet to the key-puncher who punches in the accession number, the appropriate topic (link) and role, after which the cards are ready for processing.

Indexing in depth directly from the reports is slower than indexing from the report announcement, requiring from 30–40 minutes per report and averaging 23.2 term-roles per report. The average time has been two reports per hour, equivalent to 0.5 hour per report and 0.0215 hour per locator. Since higher paid technical employees are required, the costs average $2.25 per report indexed and $0.097 per term-role used.

MACHINE ASSISTANCE

Key-punching for 9,020 reports created 155,295 cards including title, authors, and job number in addition to the terms. This was an average of 17.2 cards per report and cost $4,577, equivalent to $29.47 per thousand cards or $0.51 per report. The first batch contained 1.7% errors but this was reduced with experience down to 1.0% errors at the end. Proofreading and correction of errors required 0.18 hour per report and added a cost of $0.86 per report. The corrected cards will be transferred to magnetic tape from which will issue the following outputs:

1. A bibliographic abstract showing title, authors, report and job numbers and a list of locators arranged in a logical sequence according to roles with the role codes interpreted in a stylized form (Fig. 7).
2. An alphabetical list of authors with appropriate references by accession numbers for each.
3. A list of report numbers by divisions with appropriate accession numbers associated with each.
4. A list of job numbers by divisions with appropriate accession numbers associated with each.
5. A coordinate index in dual dictionary form, in which locators are alphabetically arranged and divided by role codes. This will include an empirical chemical formula index with an arrangement of appropriate accession numbers under each for easy manual searching.

The completed five-part index will be distributed at group leader level to the research staff for ready reference as ideas may occur at the work bench. Copies of the pertinent reports are requested from the Technical Reports Librarian.

It is planned to up-date the bibliographic abstracts weekly by issuing the weekly accumulated accessions. Supplementary searching indexes by authors, report numbers, job numbers and locators will be issued quarterly, cumulative through the year with complete revision of the indexes annually.

```
  23502                                          23502

PETROLEUM ADDITIVES SALES SERVICE - 1960
GODT H C
MERTEN H L
GRAHAM G R
HEIMSCH P F
KILLEBREW R L
REPORT NO.  2701  TYPE F  SOURCE NR 002
JOB NO. N  4049              YR. 61 MO.  2 DAY  1
  TOPIC  A
STUDY OF          SYNTHESIS
STUDY OF          TENTATIVE PROCESSES
STUDY OF          SPECIFICATIONS
PRODUCT, OUTPUT   ADDITIVES
USING, USED FOR   OIL
                  YR-1960


  23503                                          23503

EXPLORATORY INVESTIGATIONS - PREPARATION OF
  BIPHENYL-4-CARBOXYLIC ACID
ELLENBURG A M
BUSCH W A
REPORT NO.  2702  TYPE F  SOURCE NR 002
JOB NO. N  3662              YR. 61 MO.  1 DAY 27
  TOPIC  A
STUDY OF          SYNTHESIS
STUDY OF          CARBOXYLATION
PRODUCT, OUTPUT   4-BIPHENYLCARBOXYLIC ACID
REACTANT          SODIUM
REACTANT          4-CHLOROBIPHENYL
REACTANT          CARBON DIOXIDE
USING, USED FOR   CARBOXYLATION
                  LABORATORY
```

Fig. 7.—Typical bibliographic abstracts.

EXTERNAL LITERATURE

It is easy to postulate that if the publishers of technical literature will generate at the source locators for each article which presents significant new knowledge as the American Institute of Chemical Engineers is now doing and as has been recommended by the Engineers Joint Council[5], and will simultaneously generate the same information as a by-product in machine processible form (*i.e.*, punched paper tape or punched cards) for sale to subscribers, each subscriber could then apply a simple conversion program and integrate the pertinent external literature with his own internal information retrieval system. There would thus be made available at low cost to the researcher the most precise selection of both internal and external information pertinent to his topic of current interest.

Acknowledgments.—"No man is an island entire of itself." Worthwhile accomplishments are the results of effort of many people, the blending of many talents. What has been recounted above has been made possible by the cooperation of Miss Margaret Madden, Central Reports Librarian; Mrs. Beverly Edwards, Chief Indexer; Miss Juanita McCarthy, Dr. Paula Raizman and Mr. Michael Dub, serving jointly as an Editorial Board in the refinement of the dictionary and the creation of a thesaurus; Miss Jeanette Livasy, programmer; Mrs. Margaret Hernandez, key puncher and keen-eyed proofreader and corrector of errors; and the counsel and advice of Mr. J. C. Costello and Miss Carol A. Penn, consultants. The whole project could not have been initiated or sustained progress without the vision, patience and energetic drive of Dr. John R. Van Wazer, Senior Scientist, in charge of Monsanto's new Information Center of which the described system is a part.

REFERENCES

(1) Costello, J. C., Jr., "Storage and Retrieval of Chemical Research and Patent Information by Links and Roles in du Pont," *Am. Document.*, **12**, 111–120 (1961).

(2) Costello, J. C., Jr., "Uniterm Indexing Principles, Problems and Solutions," *Am. Document.* **12**, 20–26 (1961).

(3) Morse, Rollin, "Information Retrieval," *Chem. Eng. Prog.*, **57**, 55–58 (1961).

(4) Holm, B. E., "Information Retrieval—a Solution," *Chem. Eng. Prog.*, **57**, 73–78 (1961).

(5) Summary of E.J.C. action plan for indexing technical articles at time of publication and preparation of a thesaurus of engineering terminology, presented at a symposium in New York City, January 17, 1962.