A MACHINE-BASED INDEX TO INTERNAL RESEARCH AND ENGINEERING REPORTS

By G. JAHODA, M. D. SCHOENGOLD and T. J. DEVLIN

Esso Research and Engineering Company, Linden, New Jersey

Esso Research and Engineering Company has been indexing its technical reports since 1938. Alphabetico-classed subject and author indexes have been issued annually in book form. Copies were placed in libraries and file rooms and also were distributed to some technical men.

The reports indexed present the work of technical men of the Esso Research and Engineering Company and other affiliates of the Standard Oil Company (N.J.). The reports collection is now growing at a rate of about 1600 items per year.

About three years ago, Esso Research initiated an intensive study of indexing systems and index usage by its technical personnel to determine what improvements were desirable in this indexing program. This paper will summarize the results of this survey and report on the development and present status of a machine-based coordinate index to internal reports.

EXAMINATION OF COMPANY NEEDS

A random-sample interview survey of 400 technical men in Esso Research and major affiliates of Standard Oil Company (N.J.) brought out the fact that the alphabetico-classed index to internal reports was used very little. Reasons were not cited, but it appears likely that the index was not used substantially chiefly because the average technical man found it difficult. This may have been because of the complexity resulting from attempts to provide various approaches to the information.

In the next step of the Esso Research study, actual rather than hypothetical approaches to reported information were determined. An analysis was made of 281 questions for information from technical reports to determine the depth, specificity, and types of indexes which would best lead to answers to these questions. The results and some of the implications of this study were used to guide the development of a new index.

About two-thirds of these questions could be answered with a relatively shallow index, namely, an index to titles and to major parts of the reports. The rest of the questions required an index of greater depth, often to information in a single paragraph. This represented a sufficiently large portion of questions to warrant construction of a deep index.

About three-fourths of the questions were specific; that is, specific terms or concepts were required as access points to the index. Since specific access points are best provided by an alphabetic subject index (ASI), an experimental ASI designed chiefly to answer specific questions was prepared to about 150 reports, and was tested by some of the research and engineering staff. Generic relationships were minimized, so that the index was relatively easy to prepare and to use. Reaction from the test audience was favorable. A decision therefore was made to include an alphabetic subject index in the program.

About one-half of the remaining questions could best be answered with a classified index; the other half with a coordinate index. A decision to prepare a coordinate index was made for the following reasons: (1) it is easier to build a coordinate index which will answer questions intended for a classified index than vice versa; (2) the publication of a simple alphabetic subject index intended only for answering specific questions is made possible by supplementing this ASI with a coordinate index; and (3) it is anticipated that the percentage of questions which will be answered with the coordinate index will greatly increase when users become acquainted with its possibilities.

THE ESSO RESEARCH SYSTEM

The Esso Research system is based largely on previous developments in conventional library techniques and coordinate indexing. Rules for indexing, "see," "see also," and "see from" references are standard library techniques. They were used as early as 1898 in the American Library Association's list of subject headings for use in dictionary catalogs.[1] Calvin Mooers' work was consulted as regards the selection and categorization of major descriptors.[2] The concept of minor descriptors was suggested by the work of the Western Reserve University group on differentiating concepts with the same semantic factors.[3] Role indicators and descriptor links were suggested by the work of the United States Patent Office.[4,5]

The system is presently designed for use with the IBM 101 Electronic Statistical Sorter, but with provision for translation to more sophisticated systems when desirable. The code used will be touched on in the appropriate parts of the following description of system components.

Rules for Indexing. — Rules for indexing for both the coordinate index and the alphabetic subject index have been formalized and included in a short manual. This manual includes instructions on what parts of the report to read for indexing, what information to select, how to translate this information into indexing language, and what records to keep for the index.

Major Descriptors. — Major descriptors are the words or phrases which constitute the basic building blocks of the indexing vocabulary. Fewer than 400 major descriptors are currently used

in the Esso Research system to characterize the most important aspects of petroleum-refining technology.

Major descriptors are grouped into eleven mutually nonexclusive categories. The selection of major descriptors was and still is somewhat an evolutionary process. Terms were selected from such tools as the previous alphabetico-classed subject index, glossaries, and reference books. Subject experts were also consulted.

The list of indexing terms so compiled was tested on a sample of 1000 documents. The frequency of the use of each descriptor was determined and descriptors used in the characterization of more than 5% of the documents were further broken down whenever possible. For example, the descriptor "hydroxyl compounds" was broken down into three descriptors: mono, di, and poly (more than two) hydroxyl compounds. This process was repeated several times in order to refine the indexing vocabulary. An excerpt from the list of major descriptors is given in Fig. 1.

Fig. 1.– A Portion of the List of Major Descriptors

Uses of Products or Equipment

Engine
Diesel-59/4
Free piston-60/4
Gas-61/4
Gasoline-62/4
⋮

Equipment
Blower-27/5
Boiler-47/6
⋮

Parts
Baffle-24/5
Bearing-26/5
⋮

Other Uses
Coating-10/5
Laminating-11/5
Construction material-62/6
Metal working-12/5
⋮

Modifiers
Agriculture-39/5
Automotive-40/5
Aviation-41/5
Home-32/6
Industrial-42/5
Marine-43/5
Rail-44/5
Refinery-45/5
Rocket-46/5

Organic chemicals are characterized by substructure, that is, by the arrangement of the carbon chain, the number of carbon atoms, and functional groups. Hydrocarbons are also characterized by the number of side-chains. An example of the characterization of an organic chemical is given in Fig. 2.

Fig. 2.– Characterization of an Organic Chemical by Substructure

n-Pentane

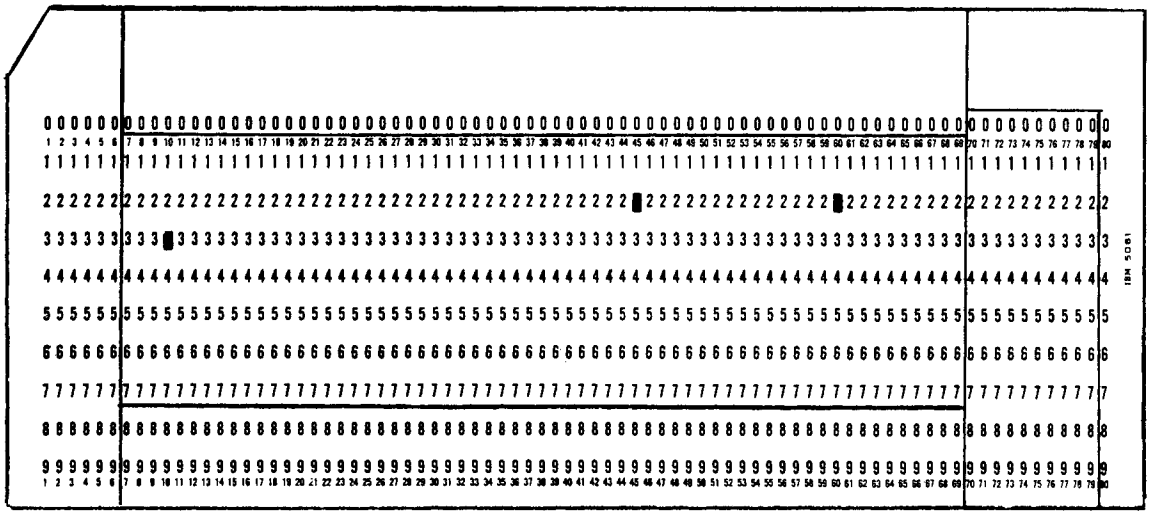
n-Paraffin
C_5 compound
Hydrocarbon

A special code for chemicals was developed after a review of existing notation systems. None of the existing notation systems met the requirements of this indexing which may be summarized as: (1) space on IBM card for encoding three or more chemicals by substructure; (2) space on IBM card for encoding concepts other than chemicals; (3) use of column-by-column sorting equipment.

This technique does not characterize every chemical to the exclusion of every other chemical. However, tests of the index indicate that this arrangement is sufficiently discriminating for our purpose.

Major descriptors are translated into single punches on the IBM card, in rows 1 to 7 of columns 7 to 69. This allows a maximum of 441 positions, each standing for a single major descriptor. In Fig. 3, n-pentane is encoded as 45/2 (the "2" punch in column 45) for n-paraffin, 60/2 for compounds with 5 carbon atoms, and 10/3 for hydrocarbons.

Minor Descriptors. — Minor descriptors are used when a concept cannot be expressed adequately by one or more major descriptors.



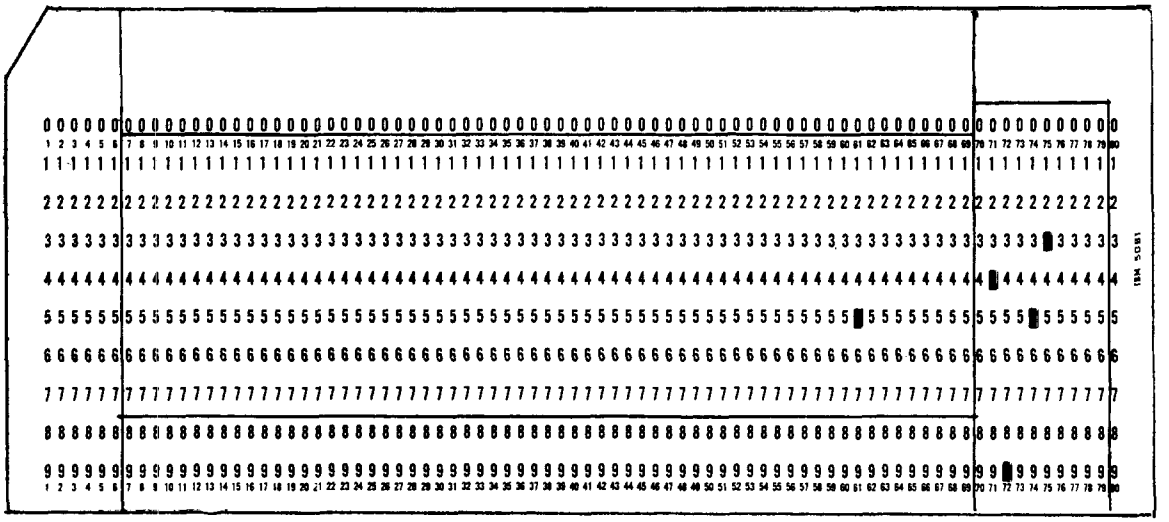
n-pentane

Fig. 3.– Encoded Major Descriptor

These minor descriptors are generally very specific. They constitute the open-end part of the descriptor vocabulary.

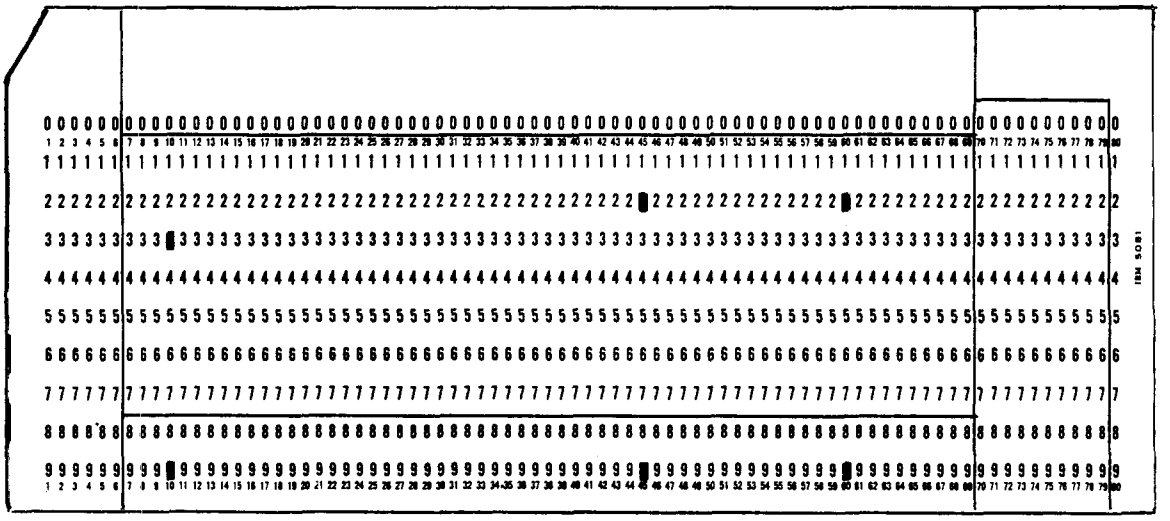
Control of minor descriptors is obtained by also assigning to each one a more generic major descriptor. For example, the minor descriptors "Esso low temperature isomerization" and "Tia Juana crude" have "Isomerization" and "Crude" as their respective major descriptors. Major descriptors for some minor descriptors have sufficient information content to permit searches on two levels. For example, both "Esso low temperature isomerization" and "Isomerization" are meaningful search terms when used in conjunction with other descriptors and in some cases when used by themselves. However, other minor descriptors are assigned to major descriptors which have no retrieval utility when used by themselves. Examples of such major descriptors are "Equipment, Other," "Process, Other," and "Functional Group, Other."

Minor descriptors are encoded as four punches in a 100-position random superimposed field (columns 70-79). The corresponding major descriptor is encoded as a single punch in the major descriptor field. Thus the code for "Esso low temperature isomerization" shown in Fig. 4 consists of a punch in column 61 for the major descriptor "Isomerization" and the appropriate punches in columns 71, 72, 74, and 75 for the specific minor descriptor.



Esso low temperature isomerization

Fig. 4. – Encoded Minor Descriptor



n-pentane as reagent

Fig. 5. – Encoded Major Descriptor with Role Indicator

Up to twelve minor descriptors can be encoded in the 100-position random superimposed code field of a single IBM card before interaction among codes results in an appreciable number of false drops. This was pointed out by Wise[6] in a discussion of a 100-position random superimposed code without an additional direct code.

Role Indicators. — Role indicators are used to describe the part or role that a chemical plays in a reaction. Three such roles are specified: Reagent (reaction with), Intermediate,

Final Product (formation of). Other roles are specified through the use of major descriptors, such as "catalyst," "support," and "solvent."

Role indicators permit one to search for information on two levels; namely, searches for chemicals as reagents, intermediates, or final products, and searches for chemicals with roles not specified. Another use of the role indicator is its combination with a chemical to characterize a process. For example, esterification is characterized by the descriptor for ester and the role indicator "formation of." The use of role indicators is illustrated in Fig. 5. The role indicator "reagent" is encoded as a "9" punch in all the columns in which the chemical is encoded, namely, columns 10, 45 and 60.

Descriptor Links. — Descriptor links are used to connect related descriptors in an encoded report. This is done to reduce false drops caused by interaction among unrelated descriptors. A coordinate index without any false drops would probably require an unjustifiably high input cost. Assume that a report deals with n-hexane and aniline. These chemicals are characterized by descriptors listed in Fig. 6.

Fig. 6. – Unlinked Chemicals

(1) n-Hexane	(2) Aniline
n-Paraffin-45/2	Aromatic-36/2
C_6-61/2	C_6-61/2
Hydrocarbon-10/3	Amine-17/3

The same combination of these descriptors is also used to characterize benzene and hexylamine, so that a search for these chemicals would result in the retrieval of this report, i.e., a false drop, unless some other step were taken.

In order to reduce the interaction among unrelated descriptors, a common "x" punch is assigned to n-hexane descriptors, and a common "y" is assigned to aniline descriptors, as shown in Fig. 7. If one then searches for linked descriptors for benzene or hexylamine, i.e., descriptors with a common x or y, one will not get the report described earlier, and will thus have avoided a false drop.

Fig. 7. – Linked Chemicals

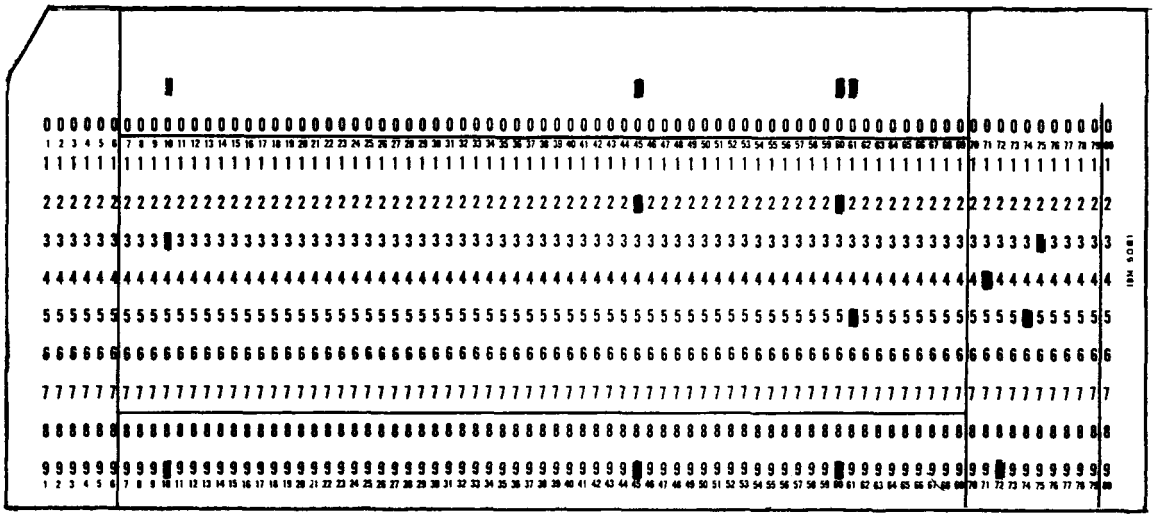
(1) n-Hexane	(2) Aniline
n-Paraffin-45/2,x	Aromatic-36/2,y
C_6-61/2,x	C_6-61/2,y
Hydrocarbon-10/3,x	Amine-17/3,y

The x, y, and 0 punches are used for descriptor links. These links are only used for descriptors in categories which are likely to cause false drops, namely, chemicals, properties, chemical modifiers, and processes. Moreover, descriptor links are used only for reports which have a sufficiently large number of descriptors to cause false drops. This is made possible by searching for the absence of descriptor links as a positive relationship.

When more than three descriptor links have to be used in coding a report, the same descriptor links can be repeated under certain circumstances. For example, the same descriptor links can be used for linking inorganic chemicals and organic chemicals, since the individual descriptors which are used to characterize these chemicals do not interact extensively. Alternatively, the report can be encoded on more than one IBM card.

An example of the linking of a chemical to a process is given in Fig. 8. Such linking does not eliminate false drops, but reduces them to what we believe will be an acceptable level.

The search for linked descriptors with a column-by-column IBM sorter is a novel feature

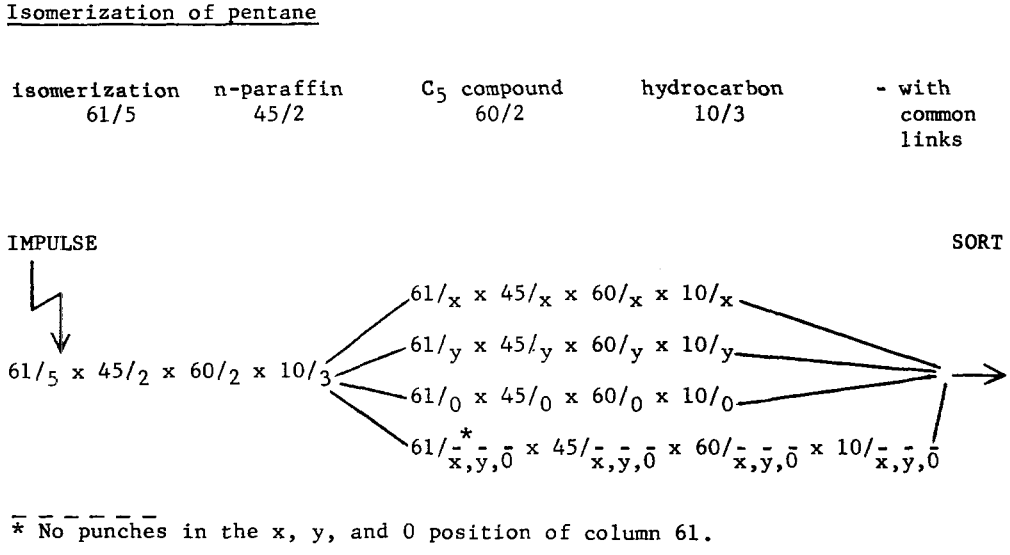


n-pentane as a reagent linked to
Esso low temperature isomerization

Fig. 8. – Encoded Linked Major Descriptors for Chemical with Role Indicator and Minor Descriptor

of the Esso Research system. The logic for the machine program for searching linked descriptors is given in Fig. 9.

Subject-Authority List. — For a system such as Esso Research's, a subject-authority list is essential for consistency in indexing. The



* No punches in the x, y, and 0 position of column 61.

Fig. 9. — Logic of Panel Board Wiring for Selecting Cards with Linked Descriptors

subject-authority list is in essence a record of all indexing decisions. Included are definitions of terms, cross-references for synonyms, related terms, and uses of combinations of descriptors. The three basic types of entries are shown in Fig. 10.

The entry for the major descriptor includes its official abbreviation, its code, its full name, a definition with a citation to an internal report or dictionary, and "see also" (s.a.) and "see from" (x) cross-references. Underlined "see from" references are terms which are included under the major descriptor as separate minor descriptors. Non-underlined "see from"

VISCOS 26/2
Viscosity
Resistance of a liquid to flow. The best known viscosity units are: Saybolt Universal, Redwood, Engler, and Absolute (SDD)

s.a. CONSIST
x Intrinsic viscosity
x Mooney viscosity
x Rheology

Mooney viscosity 26/2-70/8-71/1-72/8-75/6
VISCOS
Primarily a measure of processability of rubbers, i.e., polymer breakdown and attendant viscosity loss during processing (CRD.1UL.59)

Biodegradation USE BACT-12/6
DECOMP-56/5

Fig. 10. — Subject Authority List Entries

references are terms which are included under the definition of major descriptor, without a separate minor descriptor.

The entry for a minor descriptor includes a documented definition, whenever necessary, the code, and cross-references. The major descriptor is identified in the code.

The entry for cross-references includes the descriptor or descriptors with codes.

The subject authority list is kept in loose-leaf notebooks. The original was prepared on a Flexowriter so that a punched paper tape was a by-product. When pages are revised, additions or corrections are made on the Flexowriter tape for the appropriate pages, and the corrected tape is then used to prepare copies of the revised pages.

PRESENT STATUS

A considerable amount of time and effort has been spent in developing the system, testing it for shortcomings, and correcting these shortcomings. Generally speaking, it has been much easier to find shortcomings than to correct them. For example, it did not take long to determine that some descriptors would be used too frequently to have sufficient discriminating value, but it took considerably longer to find substitute descriptors with the proper degree of discrimination and to incorporate these descriptors in the index.

Over 3000 published and internal documents were indexed during the development of the system. After each major change in either the descriptor vocabulary or in the rules for indexing, 500 to 1000 documents were indexed, and the index to these documents was tested with a set of test questions. Some of the machine searches were checked by manually searching files of document abstracts.

Results from the most recent test searches indicate that a retrieval completeness of close to 90% can be achieved for the average search. For searches which require completeness of retrieval nearer 100%, more generic searches than called for in the question have to be made.

This results in more manual weeding of the selected documents, but this penalty is willingly accepted as the price of a simple, inexpensive system.

Actually, false drops in test searches of an index to 1000 published documents have averaged less than 10%. The index to internal reports is not yet large enough to yield a measure of false drops for the report collection.

The Esso Research system thus appears to work satisfactorily on a small scale. This is encouraging, but is not the final answer; most systems work satisfactorily on a small scale. A full-scale experiment is now being conducted to give the system a more rigorous test.

In April 1960 a two-week indexing course was given to four part-time indexers. During the training program, each indexer worked independently on the same reports in order to determine whether an acceptable consistency of indexing can be achieved with the system. Complete uniformity of indexing was not achieved, but was not expected. However, good agreement did exist in the selection of the key descriptors for each report. Moreover, since all indexing decisions are recorded in the subject authority list, recorded differences of opinions can be eliminated by the index editor. Indexing of current and past reports under the new system is underway.

This paper has attempted to summarize the Esso Research work in the field of machine-based coordinate indexes. In essence, the Esso Research system is a blend of selected features of conventional and coordinate-indexing systems, with the control of the indexing vocabulary and rules for indexing as the two basic ingredients. Future papers will report on the use experiments described and on details of the system if these experiments prove successful.

REFERENCES

[1] American Library Association, "List of Subject Headings for Use in Dictionary Catalogs," 2nd ed., American Library Association, Boston, 1898.

[2] Calvin N. Mooers, "Zatocoding and Developments in Information Retrieval," ASLIB Proceedings, 8, 3-22 (1956).

[3] J. W. Perry and Allen Kent, "Tools for Machine Literature Searching," Interscience Publishers, Inc., New York, N. Y., 1958, p. 241.

[4] B. E. Lanham, J. Leibowitz and H. R. Koller, "Advances in Mechanization of Patent Searching—Chemical Field," Preliminary Report, Department of Commerce, Washington, D. C., 1956, p. 10.

[5] D. D. Andrews, J. Frome, H. R. Koller, J. Leibowitz, and H. Pfeffer, "Recent Advances in Patent Office Searching: Steroid Compounds and ILAS" (in Information Systems in Documentation, J. H. Shera, Allen Kent, and J. W. Perry, Interscience Publishers, Inc., New York, N. Y., 1957), p. 469-71.

[6] Carl S. Wise, "Mathematical Analysis of Coding Systems," (in R. S. Casey, J. W. Perry, Madeline M. Berry, and Allen Kent (ed.), "Punched Cards," 2nd ed., Reinhold Publishing Corp., New York, N. Y., 1958), p. 447-60.