Designing an Information Awareness and Retrieval System for Chemical Propulsion Literature*

HERMAN SKOLNIK and BENN E. CLOUSER
Hercules Incorporated, Hercules Research Center,† Wilmington, Del. 19899

Received April 2, 1970

A centralized information system for the chemical propulsion literature of interest to four major locations of Hercules Incorporated is described. The system was designed to be a timely and comprehensive awareness and retrieval mechanism for documents received at each of the four locations. On realizing this objective successfully, the centralized system replaced the four established independent systems and resulted in an appreciable cost savings. A valuable product of the computerized retrieval system was the Multiterm indexing system which was conceived and developed in response to the problems that arose in centralizing and computerizing the operations. The advantages of Multiterm indexing are discussed, viz., computer-orientation, high information content, unique clarity of communication, inherent internal consistency, and high retrieval efficiency.

Research and development programs for the Chemical Propulsion Division of Hercules Incorporated are carried out at four major locations. As these programs evolved, each location set up its own document control center and information system to aid Hercules scientists and engineers at each location in solving their information problems. The information groups at each location maintained good communications and cooperation with each other. Each, however, developed its own information system from document accession to awareness and retrieval. Most of the documents handled and processed consisted of Hercules technical reports and technical reports of agencies of and contractors to the U.S. Government. Many of the Government reports are received under automatic distribution from Hercules involvement in Government contract work; others are obtained by our scanning of and ordering from Government semimonthly awareness media, such as the *Technical Abstract Bulletin* (TAB) of the Defense Documentation Center (DDC)[9], the *U.S. Government Research and Development Reports* (USGRDR) of the Clearinghouse for Federal Scientific and Technical Information,[7] and the abstract journal *Scientific and Technical Aerospace Reports* (STAR) of the National Aeronautics and Space Administration (NASA).[2, 10]

Although there was much duplication of documents automatically received or ordered, each location had in its document control center some documents not in the other centers, and the other centers were unaware of the nonduplicated documents. There was considerable duplication of effort, of course, in indexing, abstracting, and setting up awareness and retrieval mechanisms for those documents held in common. Furthermore, in one or two locations, many scientists and engineers scanned TAB and the other awareness media themselves, even though it was a time-consuming task which most scientists and engineers would rather delegate to an information system that meets their needs.

As early as 1960, two locations, the Hercules Home Office and the Hercules Research Center, which are only a few miles apart, assigned primary responsibility to the Technical Information Division at the Research Center for a chemical propulsion information system. The system was set up so that one chemist had the responsibility for scanning awareness media and ordering documents that met the requirements of the scientists and engineers of the two locations. He was also responsible for indexing and abstracting the documents received, for issuing an awareness bulletin, and for maintaining an index for retrieving the documents on request. This cooperative venture resulted in elimination of duplicate work. More important, from the viewpoint of the chemical propulsion research and development management, the awareness bulletin markedly increased the reading of essential reports by the scientists and engineers.

This success prompted the research and development management in 1966 to request a study of the feasibility of setting up a company-wide chemical propulsion information system. Although cost savings was one of the objectives, the major objective was to have an information system that informed scientists and engineers promptly and meaningfully of the evolving report literature pertinent to their research and development assignments. The study led to the system described in this paper.

ESTABLISHING NEEDS OF USERS IN FOUR LOCATIONS

In replacing four decentralized information systems with a centralized system, it is not enough to design one that looks like or consists of the total of the four. Our approach was to examine the total needs in relation to the research and development programs as though we were starting fresh. This study was made through a committee composed of the heads of the documentation services, the manager of engineering of the Chemical Propulsion Divi-

* Presented before the Symposium on Information System Design at the Fifth Middle Atlantic Regional Meeting, Newark, Delaware, April 2, 1970.

† Hercules Research Center Contribution Number 1514.

sion, and the chemist from the Technical Information Division who would be responsible for the centralized information system. The heads of the documentation services were responsible for establishing the needs of the users and for communicating this information to the centralized service.

Inasmuch as documents are generated at each location and documents are received on automatic distribution or by ordering through one of the Government's abstract publications, accessions lists at each location constituted a good measure of areas of interests. Furthermore, we visualized a savings by replacing the separate accessions lists with a central one issued promptly and with document control numbers for each location, AD document numbers, bibliographic information, abstracts if needed, and index terms.

To keep typing input at a minimum, a card was designed (Figure 1) which accommodated the needs of each document control center and which could be used directly in preparing a centralized awareness bulletin, in document card files for each location for its own holdings, and as input to the computer for printout of indexes. This is accomplished with a single typing for each document card.

Because the weekly awareness bulletin is arranged alphabetically by source (company or agency) and three of the four locations wanted to maintain a card file arranged by source, the source and the document number assigned to the report by the source constitute the first line on the card (Figure 1). The title of the report (in capital letters), date of the report, and author or authors are entered next on the card. This is followed by an abstract, if one is necessary, and finally the Multiterm or Multiterms. The abstract number, control number at each location, and the AD or NASA accession number, if known, are placed in the column along the right side of the card. The abstract number, which is assigned on assembling the weekly bulletin, also shows the security classification of the report (e.g., unclassified, confidential, or secret), downgrading number, and distribution limitation (e.g., NOFORN).

The document card was designed to be used with aluminum panels, described previously,[5] for the preparation of the awareness bulletin via Multilith masters.

Document cards are prepared daily at each location as documents are received and assigned control numbers, and sent in batches to the Research Center. By checking against the computer-printed source index and the newly-received document cards from the other locations, duplicates are eliminated. About 35% of the accessions reported from the four locations generally are duplicates of each other or of documents already in the system. When a duplicate is found, control numbers and the AD number, if known, are added to the document record in the computer for subsequent printouts. The system yields six computer index printouts—subject, author, Government contract number, source (company or agency), abstract number (assigned in ascending order as each awareness bulletin is issued), and control numbers. The control number printout is in four parts, each beginning with the control numbers of one location and relating

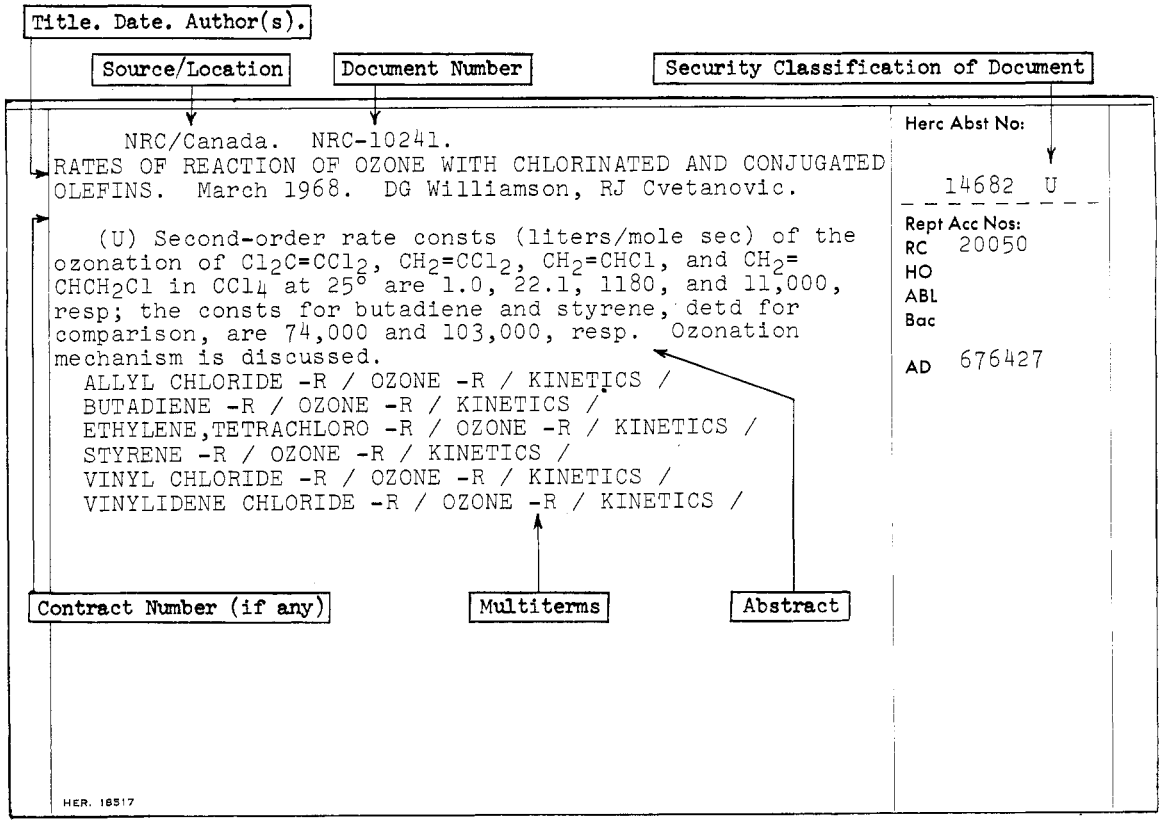

Figure 1. Card input for the chemical propulsion information system

them to the abstract number and the control numbers of the other locations.

The weekly awareness bulletin averages about 130 abstracts, and from January 1, 1967, to January 1, 1970, a total of about 17,000 documents were put into the system. As each bulletin is issued, a computer printout of a subject index to the documents in the issue is produced and sent to the document control centers. A cumulative subject printout is produced monthly. The other indexes are issued on a periodic cumulative basis.

The weekly issues of the bulletin constitute a bibliographic-abstract reference work for the chemical propulsion report literature at the four locations. Because it is arranged in ascending abstract number order, it is a very useful unit in conjunction with the six computer-produced indexes.

The subject index, however, is the heart of the system, and until it was conceived and set up, a completely computerized system was difficult to achieve.

THE MULTITERM SYSTEM

In retrospect, the Multiterm Index was conceived to give users the advantages of the generic nature of classification, the specificity of good subject indexing, the freedom of uniterm indexing, and the advantages of computer processing as described by Skolnik.[6]

A Multiterm is a series of subject headings separated by virgules (points of permutation, or wrap-around, in the computer). The subjects of the input Multiterm are arranged from generic to specific in a logically assigned order. Within a subject, a generic/specific relationship is shown with a colon, as in the following

Propellant:Hybrid
Propellant:Liquid
Propellant:Solid

A modified subject, as in the following, uses a comma between the subject and the modifier:

Stability,Burning
Stability,Light
Propellant:Solid,AP-Contg.
(AP = ammonium perchlorate)

The following letters are used to indicate action on or by, or roles, of subjects:

A = analysis	Q = property
D = determination	R = reaction
E = effect	T = treatment
P = preparation	U = use or application

Thus, the following Multiterm, which was assigned to a document as input to the computer:

PROJECTILE/AERODYNAMICS/MACH 1.5/SHAPE -E/

yields the following additional Multiterms in the computer printout:

AERODYNAMICS/MACH 1.5/SHAPE -E/PROJECTILE/
MACH 1.5/SHAPE -E/PROJECTILE/AERODYNAMICS/
SHAPE -E/PROJECTILE/AERODYNAMICS/MACH 1.5/

Under each of the four Multiterms above are listed the abstract number and the title of the document, the latter truncated to 68 characters.

Because we wanted the computer printouts on 8½ × 11-inch paper, each Multiterm is also restricted to one 68-character line. The 68-character restriction encourages the use of well accepted abbreviations and acronyms, such as BOR for BODY OF REVOLUTION and AP for AMMONIUM PERCHLORATE. When an abbreviation or acronym is used, its definition and a cross from the full term are put into the computer for printout in the subject index, for example

BOR = BODY OF REVOLUTION
BODY OF REVOLUTION,SEE BOR
AP = AMMONIUM PERCHLORATE
AMMONIUM PERCHLORATE,SEE AP

Further, the 68-character limitation motivates the indexer to consider carefully the true range of the subject content of the document being indexed, and to stay within this range with the assigned terms. In a very important sense, the Multiterm philosophy of proceeding from the most generic to the most specific terms that define the contents of a document is in harmony with the 68-character limitation. For example, the document defined by the following Multiterm,

PROJECTILE/AERODYNAMICS/MACH 1.5/SHAPE -E/

is related to another document that has the following Multiterm shifted further to the left generically:

WEAPON SYSTEM/PROJECTILE/
AERODYNAMICS/MACH 1.5/

On the other hand, the Multiterm for a document on a test method for determining the effect of configuration on projectile aerodynamics would be shifted toward the specific, such as:

PROJECTILE/AERODYNAMICS/MACH 1.5/
SHAPE -E/TEST METHOD/

Enough overlap occurs so that the user of the index will be directed from any given subject term, by more generic and more specific Multiterms, to related documents that he might not otherwise know existed. A major advantage of the Multiterm system is the ease with which a user is directed to documents that meet his needs and the elimination of documents that do not meet his needs.

Several Multiterms are assigned to documents whose contents are relatively wide-range, such as:

PROJECTILE/AERODYNAMICS/MACH 1.5/
SHAPE -E/TEST METHOD/

and

WEAPON SYSTEM/PROJECTILE/AERODYNAMICS/

In this event, at least one term is used in common.

Deviations from a consistent pattern of usage or from the needs of the users are clearly evident in the index printout, and changes can be made readily for subsequent printouts. For example, the concept MACH is a short expression for velocities relative to the speed of sound, and is made specific by adding a number or range. The use of MACH and a number or range was in direct response to user feedback when the indexer used PROJECTILE,SUPERSONIC. In the Multiterm system, we have the flexibility of using MACH <1, MACH 1, MACH >1, and MACH >5, or MACH with specific numbers,

e.g., MACH 1.5, in which MACH is a subject term related as necessary to other terms, e.g., PROJECTILE or MISSILE.

Experience revealed that the subject term became less effective as the number of entries became very large. Because the computer enabled us to insert a cross reference containing a continuously updated alphabetical list, MOTOR:SOLID, for example remains a generic subject term for documents relating to specific solid motors, without the need for it to appear in assigned Multiterms.

The logic of the Multiterm system in the chemical propulsion information system is fundamentally similar to that of a good definition, in which the general category is given first, followed by special characteristics that differentiate among members of the category.

Semantically, the Multiterm system has a high degree of clarity because of the coordination of terms that are related. For example, the subject term TANK is ambiguous, but in the following Multiterms, it is clear what tank means in the two documents:

TANK/ALUMINUM/CORROSION/
NITROGEN TETROXIDE/
TANK/GUN/AMMUNITION/PROPELLANT/

A properly constructed Multiterm generally makes sense if read forwards or backwards. For example:

ROCKET:LIQUID/NOZZLE/THROAT INSERT/
ABLATION TEST/FACILITY/

can be read backwards in conventional English as: "facility for the ablation testing of throat inserts of nozzles of rockets using liquid propellants" or forwards as "liquid propellant rocket nozzle throat insert ablation test facility." The latter is an example of the relatively common engineer's "freight-train" or chemist's "polymerized" English.

COMPARISON OF MULTITERM INDEXING WITH THAT OF DDC

That the U. S. Government is heavily involved in the indexing and abstracting of technical reports issued under Government contracts is shown by the average contents of recent issues of USGRDR, STAR, and TAB:

Publication	Contents, No. of Documents (Semimonthly)	
USGRDR	1400	4200
STAR	1700	
TAB	100	

Some documents that appear in USGRDR are repetitious of the unclassified section of STAR, so that less than 8000 documents per month are new announcements. A common indexing method is used, namely indexing under a controlled vocabulary or thesaurus[1, 3, 7] which uses single terms or descriptors and coordinated terms or descriptors with cross references. Writers of reports under Government contracts are asked to supply keywords that, in their opinion, could be used for retrieval.

Thus, the report, AD 687,881, from the University of Dayton, entitled, "Structural and Orientational Aspects of Graphitized Fibers," classified in USGRDR under *Materials: Fibers and Textiles,* was assigned the following descriptors and identifiers by DDC and keywords by the author:

Descriptors	Identifiers	Keywords
*Carbon fibers	Goniometers	Interlayer spacing
Configuration	Orientation	Orientational parameters
Crystal structure		Profile analysis
Crystallography		X-ray diffraction
Graphite		X-rays
Measurement		
Optical instruments		
X-ray diffraction analysis		

*Index Entry for USGRDR

We indexed this report with the Multiterm:

FIBER:GRAPHITE -Q/STRUCTURE -D/
TEST METHOD/X-RAY -U/

Except for X-ray, there is no agreement between the author's keywords and the DDC descriptors. This should not be surprising, because authors tend to be subjective rather than objective in assigning subject terms to their papers. An author whose objective was to apply an apparatus or to study a crystal structure is apt to overlook the possible importance to somebody else of the graphite fiber he used. It is more surprising that the indexer preferred to use the term "carbon fibers" instead. When descriptors and keywords are word oriented rather than subject oriented, the result is random dispersal of related subjects throughout the index. Also, as with any index of this type, there is the possibility of false retrieval and, from many viewpoints, of no retrieval.[5] Although the subject terms in the Multiterm above were selected to relate the document to Hercules programs, they express a relationship which makes retrieval both sure and selective from four basic viewpoints. Furthermore, the generic/specific relationship inherent in the term FIBER:GRAPHITE allows the user to browse among other FIBER:X terms if he is interested in the general field of fiber-reinforced composite materials or if he is interested in materials whose crystal structure he might study with X-rays. The Multiterm most directly gives the user sufficient information for him to determine rather positively his further interest in the document.

Titles of documents in this literature, as in others, do not always describe the contents adequately. Multiterms give considerably more information. Indeed, the informational content of a properly assigned Multiterm is equivalent to a good title and short abstract.

Figure 1 illustrates the high informational content of a Multiterm relative to even reasonably good titles. The title of the document in Figure 1, however, would have been more meaningful if it read: "...Conjugated Dienes" instead of "...Conjugated Olefins." The amended title is inherent in the Multiterms assigned. The following descriptors for this document were listed in USGRDR (December 25, 1968).

Descriptors	Identifiers
Alkenes	Ozonation
Oxidation	Chlorohydrocarbons
Halogenated hydrocarbons	
Ozone	
Reaction kinetics	
Chlorine compounds	
Molecular isomerism	
Addition reactions	
O-Heterocyclic compounds	

In terms of the information in the document, some of these descriptors, and certainly combinations of some of the descriptors, are likely to yield false retrievals. The Multiterms of Figure 1, on the other hand, are directly related to the information in the document and cannot lead to a false retrieval. Furthermore, the Multiterm tends to extend the searcher's ability to find documents of interest easily. For example, if we were interested in the reactions of ozone with any chemical, we need merely to scan the Multiterms in which OZONE -R is the first term for the other reactant term—e.g., BUTADIENE -R, or STYRENE -R, or VINYL CHLORIDE -R.

ADVANTAGES OF THE INFORMATION SYSTEM

Hercules interest in Government reports in the broad area of chemical propulsion results in the accession of about 130 documents (without duplication) per week at four locations. Centralization of an information system, at the Research Center, for the preparation of a weekly awareness bulletin and computer-produced indexes has resulted in a more timely and more comprehensive system at considerably less cost than the original four systems.

The awareness bulletin, which is distributed to 180 scientists and engineers at the four locations, has promoted an appreciable increase in the use of the Government report literature with a favorable feedback from the readers. Since it was initiated, the system has informed the readers of the total documents received at all locations, including those received at their own location with their own document control numbers. The system has reduced almost completely the need for many scientists and engineers to examine TAB, USGRDR, and STAR, as they did before the centralized system was set up.

A valuable product of the information system was the Multiterm indexing system which was conceived and developed in response to the problems that confronted us in taking on this challenging task.

ACKNOWLEDGMENTS

We acknowledge the advice and encouragement of R. Steinberger and the assistance of W. R. Payson, R. H. Petty, T. M. Norback, and W. G. Young who were involved in the early stages of setting up the information system.

REFERENCES

(1) Caponio, J. F., and T. L. Gillum, "Practical Aspects Concerning the Development and Use of ASTIA's Thesaurus in Information Retrieval," J. CHEM. DOC., 4, 5–8 (1964).
(2) Day, M. S., "The Scientific and Technical Information Program of the National Aeronautics and Space Administration," J. CHEM. DOC., 3, 226–8 (1963).
(3) Gillum, T. L., "Compiling a Technical Thesaurus," J. CHEM. DOC., 4, 29–32 (1964).
(4) Green, J. C., "The Role of the Department of Commerce," J. CHEM. DOC., 3, 223–6 (1963).
(5) Hicks, M. S., "Government-Sponsored Research Reports in Three Areas of Physical Chemistry," J. CHEM. DOC., 3, 144–8 (1963).
(6) Skolnik, H., "The Multiterm Index: A New Concept in Information Storage and Retrieval," J. CHEM. DOC., 10, 81–5 (1970).
(7) Skolnik, H., Book Review of NASA Thesaurus, J. CHEM. DOC., 8, 53 (1968).
(8) Skolnik, H., and W. R. Payson, "A New Posting Method for the Preparation of a Cumulative List," J. CHEM. DOC., 3, 21–4 (1963).
(9) Vann, J. O., "Defense Documentation Center (DDC) for Scientific and Technical Information," J. CHEM. DOC., 3, 220–2 (1963).
(10) Wente, V. A., and G. A., Young, "Selective Information Announcement Systems for a Large Community of Users," J. CHEM. DOC., 7, 142–7 (1967).

A Comparative Study of a Fragmentation vs. a Topological Coding System in Chemical Substructure Searching*

MELVIN L. SPANN[1]
Science Information Facility, Food and Drug Administration, 200 C St., S.W., Washington, D. C. 2020
DELORES D. WILLIS
Statistical Data Branch, Bureau of Medicine, Food and Drug Administration, Washington, D. C.

Received June 26, 1970

For the past six years, the Food and Drug Administration has been utilizing a fragmentation coding system for the storage and retrieval of chemical structure information pertaining to Investigational and New Drug Applications. After installation of the Chemical Abstracts Service Substructure Search System by FDA's Science Information Facility, a three-month comparative study was conducted using both systems for the retrieval of chemical compound information. This paper presents many of the observations made during this study with particular emphasis on the degree of specificity available in question phrasing and the precision in retrieving chemical compounds containing desired substructures.

Recognizing the need to automate the storage and retrieval of information in Investigational New Drug Applications (INDs) and New Drug Applications (NDAs), the former Bureau of Medicine of the Food and Drug Administration, in 1963, instituted an information retrieval system called RAPID (Retrieval of Automatically Processed Information on Drugs). This electronic accounting machine-based system was established to handle chemical, medical, and management data received by FDA in conjunction with drug applications. This paper is limited

* Presented at the Fifth Middle Atlantic Regional Meeting, ACS, Newark, Del., April 1, 1970.
[1] Present address, Food and Drug Administration, Bureau of Drugs, 5600 Fishers Lane, Rockville, Md. 20852