DIVISION OF CHEMICAL LITERATURE
AMERICAN CHEMICAL SOCIETY
ATLANTIC CITY, N. J.
SEPTEMBER 9 - 14, 1962

# PRACTICAL ASPECTS CONCERNING THE DEVELOPMENT AND USE OF ASTIA'S THESAURUS IN INFORMATION RETRIEVAL

J. F. Caponio
T. L. Gillum
Armed Services Technical Information Agency
Arlington Hall Station
Arlington 12, Virginia

## ABSTRACT

This paper presents a detailed description of the nature of ASTIA's report literature control problem and the evaluation of a mechanized system for storage and retrieval of document references pertaining to science, technology, and medicine. The application of the thesaurus concept for control of an indexing vocabulary is related and the revision of the First Edition of the Thesaurus of ASTIA Descriptors is explained. Problems arising from efforts to evaluate terminology and indicate relationships among selected descriptors are discussed in terms of the broad principles derived. Emphasis is placed on the importance of the human element in a mechanized information retrieval system. The advantages of maintaining a degree of simplicity that permits efficient utilization of the system by all types of users are set forth. The desirability of compiling various aids for the indexer and bibliographer such as displays of groups of subject-related descriptors, frequencies of descriptor usage in indexing and searching, descriptor assignment relationships, scope notes or definitions of ambiguous terms, and manual files of lightly posted descriptors is suggested. The utility of the descriptor vocabulary as a means of indexing accessioned documents, formulating search strategies and compiling subject indexes for abstract bulletins and other publications is discussed on the basis of operating experience. The importance of cooperative efforts among the major information centers with respect to vocabulary development, abstract standardization, and document indexing is considered.

## INTRODUCTION

One of the most crucial problem areas in the process of research, development, testing, and evaluation is science information communication. It is quite obvious that before the results of research can be exploited effectively they must first be put under bibliographic control and made available to everyone who can use them. As the research and development spectrum expands, with increasing allocations of resources involved, the time gap between scientific and technological discovery and reporting and application tends to widen.

In the scientific and industrial community, journal publications, meetings, and symposia have been the chief means of keeping abreast of the constantly accumulating mass of new findings elucidated by scientific research. In the Department of Defense (DOD), the largest research sponsor, responsibility for collecting, analyzing, and disseminating the results of research and development is centered primarily in the Armed Services Technical Information Agency (ASTIA).

In the years since its inception, ASTIA has accumulated a vast and comprehensive collection of more than 700,000 technical reports in virtually all fields of science, technology, and medicine. ASTIA presently

receives, abstracts, and indexes 30,000 reports each year; these are announced in a semi-monthly abstract journal, the Technical Abstract Bulletin (TAB). In the next two years the total accessions are expected to reach 300,000 reports annually, as a result of ASTIA's new acquisition policy.

ASTIA makes available to DOD contractors and military organizations full sized copies of reports and, upon request, supplies bibliographies of its holdings on any given subject. At present, requests for more than 3500 reports and 25 literature searches are received per day from among 4,000 users. These figures are expected to rise in proportion to the increase in accessions.

These increases in workloads over the past five years lead, in 1959, to the adoption of automatic data processing techniques for processing report requests and performing literature searches. As part of this conversion to machine methods, the subject heading method of subject analysis, used at ASTIA since 1947, was redesigned into a concept coordination or descriptor system.

In the evolutionary steps traversed by the document processing center of ASTIA from a highly developed manual system of subject heading control to a completely mechanized coordinate system of classification, the most significant contribution was the development of an authoritative indexing vocabulary from the ASTIA Subject Heading

List. The conversion from subject headings to descriptors and the reindexing of the 200,000 most recent were achieved as a result of ASTIA's Project MARS. In this vocabulary development, the major subject headings were divorced from their subdivisions and, in one move, the list was reduced from 70,000 combinations to about 8300 main headings. The 850 subdivisions were reduced to about 600 resulting in a vocabulary of some 9000 terms.

This draft vocabulary was then edited and purified by eliminating synonomous terms and establishing appropriate cross references. Also, many of the infrequently used terms were coalesced and included in closely related terms. These actions further reduced the number of descriptors to less than 7000. Finally, the descriptors were organized into 292 groups of logically related terms and 19 subject discipline fields designed according to a quasi heirarchical classification system. A volume containing the descriptors, their cross references, and the descriptor groups and fields was published in May 1960 as the Thesaurus of ASTIA Descriptors.

## Revision of the ASTIA Thesaurus

Despite the numerous shortcomings of the Thesaurus vocabulary, not the thesaurus concept, the conversion to a descriptor coordination approach repre-

sented a milestone in ASTIA's history. In addition to the many advantages resulting from the establishment of the coordinate retrieval concept, the system also demonstrated unlimited potentialities with respect to the production of bibliographic reference and search tools. At the time of publication, the Thesaurus was known to be incomplete and known to contain many, but not insurmountable, deficiencies that would certainly be eliminated or reduced in subsequent editions. With this realization ASTIA soon initiated a research project which had as its objective the refinement and revision of the Thesaurus of ASTIA Descriptors, First Edition.

Since the thesaurus concept as developed by ASTIA had generated such keen interest among specialists in the field of information science and because it had gained widespread acceptance throughout the DOD scientific community as well as the industrial community and the general public, ASTIA enlisted the assistance and advice of subject matter specialists and documentation experts from government, industry, and education for this revision effort. By requesting outside cooperation and active participation, it was felt that vocabulary compatibility among government and non-government operated activities engaged in information retrieval could be achieved more easily and that standardization

of related bibliographic aids could be ensured. The first product of the cooperative effort was the formulation of a statement of general guidelines to be followed and objectives to be attained in the revision process (Ref 1). Under the joint auspices of ASTIA and the Engineers Joint Council, a panel of documentalists and subject specialists was organized to undertake the actual task of revision (Ref 2).

Compliance with the guidelines entailed an evaluation of each entry in the First Edition in regard to its utility for indexing at ASTIA as determined by the frequency of its use in indexing accessioned documents and its acceptability and significance as a scientific term. A determination was made of the relationships, both generic and logical, of each term to other terms in the vocabulary. Finally, explicit statements were provided regarding the use of ambiguous terms to which was attached a special meaning that was not readily discernable.

The three criteria, utility, frequency of use, and acceptability, are obviously interrelated and should properly be considered facets of evaluation rather than distinct qualities, but are separated for purposes of discussion. Examples are given to illustrate the types of decisions that were made.

Utility for retrieval is, at best, a nebulous quality and was more often than not determined subjectively by the judgment of experience bibliographers and indexers. For example, the term "Physical properties" occurs frequently in the literature and, as a descriptor, had been used often by ASTIA indexers. However, despite any connotation by the term when it appeared in context with specific chemical compounds, materials, etc., it had proven to be almost meaningless when used to phrase queries because of the wide range of possible properties involved and was abandoned, with provision for recataloging on the basis of more specific terms.

The frequency of posting of the descriptors in the indexing of some 230,000 documents gave quantitative evidence of their utility, although these statistics alone did not prove the value of the various terms. Descriptors posted ten or fewer times were suspected to be of little use and most of these were either included under more general concepts or made the object of a coordination of two or more other descriptors. For example, "Apricots" was properly included totally by the more general term "Fruits", whereas "Platinum electrodes" now appears as an instruction to use "Platinum" and "Electrodes", a logical coordination of two descriptors already in the vocabulary.

Not all lightly posted terms were disposed of, however. The descriptor "Earthquake-resistant structures" was retained although it had been used only seven times in indexing the entire collection because there were no other descriptors that could adequately convey this concept. On the other hand, many descriptors in the vocabulary had been posted frequently but did not represent really significant concepts. One example is "Control valves", a term that had been used 126 times in indexing but upon examination was found to be merely an unfortunate choice of words that had been coined several years earlier in an effort to inject some specificity into the subject heading list. Several descriptors were clearly enough defined and consistently used within ASTIA but were not entirely acceptable or understood elsewhere. One such instance is "Electrostatic capacitance", a term that had traditionally been used in lieu of the more commonly encountered "Capacitance". This situation was reversed to help make the vocabulary more palatable to those outside ASTIA.

A major change in the format of the revised Thesaurus is the development and display of generic relationships. The intent in showing these relationships is to further delineate the concept represented

by the descriptor and to provide immediate access to terms that are considered to be more general or more specific. The generic relationships are indicated by "Generic to" and "Specific to" references and are provided only where appropriate terms existed in the vocabulary, having been subjected to the evaluation of utility previously discussed. In general, no effort was made to manufacture descriptors or to introduce terms not previously used by ASTIA merely for the sake of having symmetrical cross references. At a given descriptor entry, more general descriptors are indicated by a "Specific to" reference and more specific descriptors are indicated by a "Generic to" reference. For instance, "Halogens" was listed as "Generic to" references "Bromine", "Chlorine", etc., and each specific halogen was the descriptor "Halogens" as a "Specific to" reference.

The procedure used in the First Edition of displaying descriptors considered to be logically related to a given descriptor as "Also see" references was retained. This device is valuable in the use of the Thesaurus, but the subjective judgments involved in determining useful logical relationships among descriptors present many difficulties and make virtually impossible the development of complete references. All logical relationships cannot be foreseen at the time a descriptor is evaluated, so only the more obvious cross refer-

ences were made. As a result some of the references may prove worthless, whereas there will be a need for adding references as experience dictates.

The most significant change and the one that will probably be of most benefit to users of the Thesaurus is the provision of scope notes to delineate the meaning of ambiguous descriptors and to distinguish among descriptors having partially overlapping meanings. Most scope notes take the form of explanatory phrases, with examples where necessary to convey a meaning. These notes are intended to be read as part of the cross references and reflect the use of the term by ASTIA. The term, cryogenics, for example, is often used and generally understood, but it has come to have different meanings in the fields of rocket propellants and solid state physics. Equipment for handling liquefied gases might be completely unsuitable for studying superconductivity, whereas writers in both areas often describe their studies as "cryogenics". If such a term is available in an indexing vocabulary the natural inclination of the indexers will be to index documents in both fields with the same term, creating an unreliable retrieval situation. Therefore, the descriptor was given a scope note imposing an arbitrary parameter on the temperature ranges as dictated by ASTIA usage.

No effort was made to provide dictionary definitions for terms that are unambiguously used in the literature. Terms with more than one popular meaning or with meanings that differ appreciably in various scientific disciplines were assigned the connotation imposed by ASTIA, or if used in differing connotations, were made separate descriptors by means of parenthetical expressions.

Terms that represent concepts within the scope of a given descriptor but more specific than necessary for indexing the ASTIA collection are entered as "Includes" references at the descriptor, as are popular synonyms. Such terms also appear in their alphabetic sequence with the instruction "Use" the appropriate descriptor. Hence, the descriptor "Dissociation" has as an "Includes" reference "Thermal dissociation". The latter, in its alphabetical sequence bears the instruction "Use Dissociation". Terms that represent concepts that are not expressed directly by descriptors but are indexed by a combination of descriptors also appear in their alphabetic order with an instruction to "Use" the appropriate combination. For example, "Anticorrosive points" appears as an instruction to use "Corrosion inhibition" and "Points". These references permit access to the indexing vocabulary on the basis of a recognized terminology not used

for indexing. (Representative Thesaurus entries are shown in Figure 1.)

Indexing at ASTIA is performed by professional personnel who consistently handle reports in fairly narrow subject fields. The same individuals analyze queries in their fields and formulate search patterns to locate the desired information in the collection. The general approach is to coordinate two groups of descriptors, the result being references to all documents in the collection that have in common any descriptor pair from the two groups. Several variations in search strategy are possible. All holdings posted with a given descriptor or descriptors can be retrieved without coordination; three or four groups, or levels, of coordination can be used; or document references posted with chosen descriptors can be excluded from the results of a search. Any number of descriptors can be used in a coordination. Moreover, where three or four search levels are specified, the results of each level of coordination are recorded and can be obtained as output should the complete search yield unsatisfactory results.

Machine output is in the form of punched cards containing document accession numbers. The appropriate catalog cards, complete with the bibliographic

information, abstract, and indexing terms associated with the document are mailed to the customer who may then request copies of the desired documents.

Computer techniques provide a variety of possibilities for compiling aids to the indexer and searcher and to the evaluator of the vocabulary. The frequency of use of each descriptor in indexing and searching provides some indication of the volume of output to be expected from a given search pattern. Periodic checks of these figures point out the subject areas of most interest in the collection and may reveal descriptors that should be redefined or abandoned. Tabulations of word associations in input and output give good indications of promising coordinations for searching and may indicate the need to exclude descriptors. Such associations can also indicate overlapping meanings and the need to redefine or combine terms.

Technical Abstract Bulletin Index

The development area of greatest long term potential is ASTIA's efforts directed toward more effective indexing methods and new techniques for the mechanized handling of index entries for published indexes. Because the printed index still represents an extremely efficient tool for conducting both current

awareness and retrospective searches of the literature, the automatic or "slave" print out constitutes one of the most advanced applications of machine methods in the handling of technical information.

After several months of experimentation and pilot study designed to explore methods of refining ASTIA's announcement media, ASTIA inaugurated a new subject index to TAB (Fig. 2). This new index, generated by a coordinate system of indexing provides two principal advantages. The first is that it retains the convenience of the conventional coordinate index, facilitating rapid scanability and providing the multi-aspect approach. The second principal advantage is that it minimizes the need to refer to the complete abstract entry since each index entry usually contains sufficient information to enable the user to determine with reasonable certainty whether a particular document should be read. Despite the attractive possibility of indexing and abstracting by machine which has catalyzed present day research efforts on methods of eliminating or reducing human subject analysis of documents, the new ASTIA subject index is a product of human analysis whereby each index entry is carefully formulated in accordance with systematically developed guidelines for indexing. Each entry of the subject index contains

an annotation or indicative abstract. A typical index entry is composed of the following elements: the primary descriptor, indicated by an asterisk; the annotation, a synthesized statement reflecting the subject content of the document; the division of TAB in which it appears; and the ASTIA document number.

With the appearance of the new subject index, the Technical Abstract Bulletin now provides the researcher with an effective decision making instrument and information announcement publication. The new index entry is similar to the "lengthy hybridized index entry" developed and used by Welt (Ref 3) in the Cardiovascular Literature Project of the National Research Council.

## Summary

The approach to multiple access retrieval at ASTIA has been that of formulating a dynamic, but well-controlled vocabulary of defined descriptors that are displayed, together with their interrelationships and notes to indicate their usage, in the Thesaurus of ASTIA Descriptors. The utility of this concept as a tool for storing and retrieving technical report literature and for formulating accession indexes has been demonstrated by successful use at ASTIA over a period of many months.

The present status of vocabulary revision and index preparation reflects attempts to refine a retrieval system while maintaining maximum productivity, preserving as much as possible the validity of previous work, and retaining a high degree of continuity in the system. Such a situation makes major policy changes prohibitive, but leaves opportunity for technical improvements, many of which have been incorporated into the revised Thesaurus and the new TAB index format. Even these improvements must be viewed more as temporary expedients than as long-range solutions to information problems.

Experience to date has clearly demonstrated the need for research on vocabulary and indexing problems in large, multidiscipline libraries. Many concepts of information storage and retrieval that have been utilized for smaller, more specialized collections cannot readily be employed at ASTIA or other large centers. The vast amounts of raw data accumulated by indexing and retrieval efforts at such centers must be analyzed and reduced to forms useful to documentalists.

Hopefully, future progress in the vocabulary and index development will be marked by planned programs of cooperative effort among the major information centers. This effort is of the highest impor-

tance and significance, for without it there is little prospect of bringing about the needed rapid advances in the art and science of information communication.

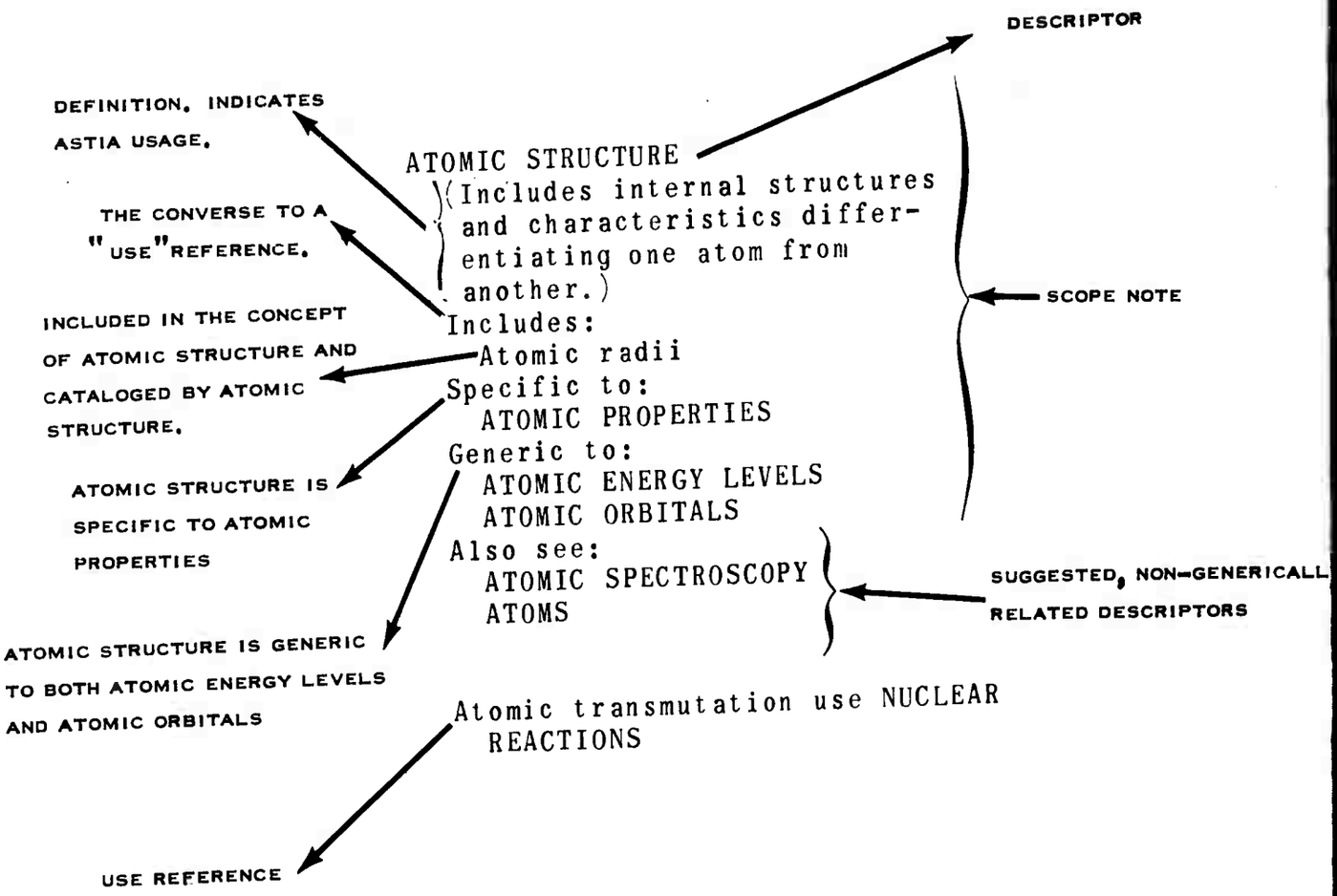
## REFERENCES


(1) T.L. Gillum, P. H. Klingbiel, C. N. Mooers, and Eugene Wall, "The Philosophy and Guidelines for the Revision of the Thesaurus of ASTIA Descriptors", November 1961.

(2) I. D. Welt, "A combined indexing-abstracting System." International Conference on Scientific Information, Washington, D. C., 16-21 November 1958. Washington National Academy of Sciences--National Research Council.

(3) Eugene Wall, Final Report on the Revision of the Thesaurus of ASTIA Descriptors. Final Rept. AF 18(600)1944, 6 August 1962, AD-278 168.

# REPRESENTATIVE THESAURUS ENTRIES



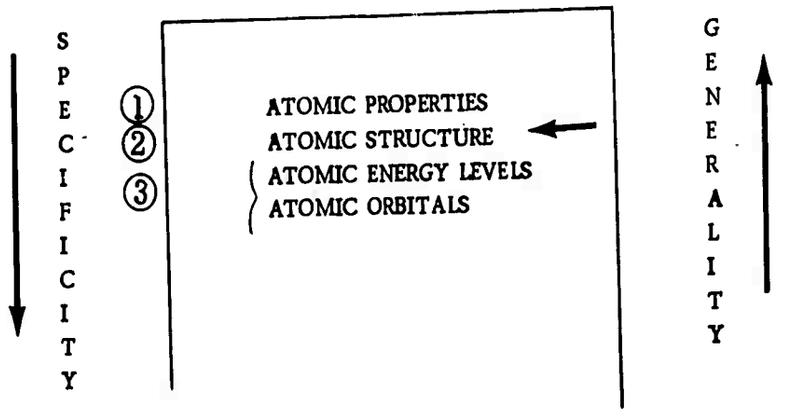




Fig. 1

# REPRESENTATIVE TAB INDEX ENTRIES

- LI
ALS WERE
INTERMEDIATE
AD-275 618

N

*CHEMICAL ANALYSIS

PREGNATED HEIR FUNC-
ZIRCONIA - MAGNESIA REFR
SILICON MIXTURES - CHEMICAL
AD-275 506 DIV.

CHROMATOGRAPHIC, IR SPECT
BOMB, AND ELECTRICAL CONDUCTI
TERMINING WATER CONTENT OF N2O4
DIMETHYLHYDRAZINE, AND HYDRAZINE
NG MARINE ELECTRICAL CONDUCTIVITY WAS MOST PROMIS.
AFT.
AD-275 537 DIV. 10

*CHEMICAL ELEMENTS

NOISE SUPRES-
TABLES OF MASS SPECTROGRAPHIC LINES OF CHEMICAL ELEMENTS RECORDED FROM THE LINES SEEN BY RF SPARK IONIZATION OF SELECTED INORGANIC SOLIDS.
AD-275 468 DIV. 20

*CHEMICAL PROJECTILES

CAVITY
S AND IN-
EQUATIONS.
ANTIPERSONNEL AMMUNITION; FRAGMENTATION AMMUNITION; FIN-STABILIZED AMMUNITION; CANISTER PROJECTILES; CHEMICAL PROJECTILES.
AD-329 355 DIV. 22A

ICATION OF
ND ELECTRON
GNETIC RELAXA-
STEMS.
*CHEMICAL REACTIONS

THE APPLICATION OF EQUATIONS TO COMPLEX REACTIONS IN EXHAUST GASES FOR USE IN DIGITA COMPUTERS. THE FLOW OF THE GAS UNDERGOING REACTIONS IN THE DIVERGENT PORTION OF A N IS CALCULATED.
AD-275 464 DIV. 9

DIATION
ED BY
A STREAMTUBE APPROXIMATION FOR OF THE CHEMICAL REACTIONS AND ION. THE INVISCID FLOW FIELD OF HYPER
AD-275 550 D

Fig. 2

UNCLASSIFIED