A Natural Language Storage and Retrieval (ABC) Method: Its Rationale, Operation, and Further Development Program*†

BERTHOLD ALTMANN
Harry Diamond Laboratories, Washington, D. C.

Received May 26,1966

The ABC system briefly described in this paper was developed to provide a measure of flexibility in solving storage and retrieval problems; in designing, performing, automating, and controlling other library activities; and in facilitating cooperation with different information activities. The initially manual system permits the information-seeking scientist to interact in scientific language with the analyzed contents of the collection. Brief indicative or informative statements, standardized for terminology and syntax, are machine processed to produce the required retrieval tools, the catalogs, the indexes, and keyword-type lists. The index allows for browsing and permits multiple approaches to general and specific subject areas as well as to quantitative data (parameters).

This paper describes a storage and retrieval system recently developed and put into operation by the Harry Diamond Laboratories, Washington, D. C. Particular emphasis is placed on the more important reasons that convinced us of the need to seek new approaches and new solutions to services that information centers operating under accepted systems can and do provide. The arguments presented summarize the extensive deliberations, surveys, and studies we were involved in during the early days in 1960 when the program was initiated.

When asked to construct an efficient reference service, we started with an analysis of the actual information requirements of our professional personnel. In this study we arrived at the following conclusions:

Because the research and development efforts in our laboratories are concentrated upon the production of new electronic and electromechanical systems, subsystems, and devices, the information office had to be mission-oriented Scientists and engineers expressed their desire to obtain information limited to their particular tasks and interests, and only to receive publications that offer definite advances in the state of the art. Thus, in order to eliminate wasted effort in acquisition, processing, and retrieval, and delays in our services, we had to subject all materials to a thorough evaluation with respect to pertinency as well as quality before integration into the collection.

The second requirement presented a still greater challenge to the system designer. Our scientists and engineers wished to play an active part in the retrieval process. They demanded "browsability," a system that permitted them to search and select without having to deal with intermediaries.

When we endorsed this second requirement, we recognized that two additional requirements were implied: First, we had to make available an understandable, visible index of the information in the collection. We had to discard the use of artificial languages (or algorithms); we had to rely exclusively upon the language in which scientists and engineers customarily communicate—that is, scientific and technical English. Second, to enable the scientist and the engineer to search for information independently of library personnel, we had to build signposts into the indexing system to guide them to all the disciplines and types of information that are related or applicable to their problems, no matter how limited or deficient the first definition of a request might be.

In addition to user requirements, we had to pay attention to the requirements of the information office itself. Of these I shall discuss only two:

The information had to be incorporated into the collection for approaches from different viewpoints, because the projects in our laboratories are frequently handled by teams of physicists, chemists, engineers, etc., all working on the solution of one common problem. This implies not only differentiation in emphasis and terminology, but also the setting apart of papers written for subject specialists from those addressing a more general population of readers.

The computer facility of the installation (a 1410–7090/94 combination) was accessible to the information office only over weekends and during some late-shift hours. There existed neither communication links nor arrangements for time sharing and real-time operations. Because our computer facility is about one mile from our location, and because we wanted to satisfy numerous short-term

* Presented before the Division of Chemical Literature, Symposium on Problems of Small Information Groups, 151st National Meeting of the American Chemical Society, Pittsburgh, Pa., March 25, 1966.

†The Army Research Office, Scientific and Technical Information Division, Washington, D. C. supports the development of the system as a contribution to the Army Technical Library Studies.

requirements of our users, we were prevented from relying exclusively upon automatic retrieval operations.

We surveyed various available coordinate indexing systems in the light of these requirements, and recognized that using any of them in our installation would have the following disadvantages.

It would not provide the scientist with "browsability," and therefore not with direct access to information. It would create unavoidable misunderstandings when documentalists and programmers had to be introduced as links between request and retrieval. It would also preclude the librarian from ever being certain of the thoroughness or completeness of a given retrieval run, because he would have at his disposal merely a list of words and not a record of the concepts covered by his collection. Finally, delays could easily occur in obtaining urgently needed information.

When we looked at the economic side of the problem, preliminary cost estimates indicated that the cost of operating these systems would equal and in most cases exceed the cost of the system we decided to introduce.

Our system, known as Approach by Concept (ABC) System (*1*), was designed to fill both the users' and the information office's requirements.

Evaluation of publications with respect to pertinence and quality;

Description of the selected documents or essays in verbless sentences or strings of phrases;

Standardization of these sentences with respect to their language building blocks, nouns, adjectives, participles, gerunds, prepositions, conjunctions, and syntax; and

Printout of a Key-Word-In-Context type dictionary (Figure 1).

Each line representing a descriptor sentence is identified by a three or four digit alphabetic accession code and by nouns or nounphrases marked by an asterisk.

A four digit code provides the capability of distinguishing about 457,000 different descriptions; an eight-letter code would provide a capability of 209 billion.

The retrieval can be performed automatically or manually after consultation of the ABC Dictionary. If a computer output is preferred, the machine can be programmed to print out all the titles filed under the selected descriptors or their respective codes.

If the information is needed immediately, the user scans the dictionary, locates the sentences dealing with aspects of his problem, and continues searching under co-occurring terms, the signposts that lead to additional clusters of descriptor sentences (this at successive stages), until his information requirements are satisfied.

Only after the system had been placed in operation did we appreciate its versatility. We observed how the investigator enriched his vocabulary, redefined his query, and adjusted and developed his search strategy continuously during the advancing process of searching. It was not the system's capability of associating terms, but its power of associating concepts, sentences, and ideas, and of establishing instantaneous relationships between different disciplines that astounded its users as well as its designers.

The retrieval of the investigator is completed in the following manner: He records the asterisk term and letter code for each selected descriptor phrase. With this information he approaches a card or book catalog arranged alphabetically by asterisk terms and codes where he locates the titles of the related publications very rapidly (Figure 2).

Space permits only this oversimplified account of the system, and does not permit an explanation of how the descriptor-phrases are standardized, or descriptions of the programs that generate, from one single input, all the required bibliographic tools, such as the dictionaries of the descriptor-sentences, the different sets of catalog cards in our catalogs, the accession bulletins and special bibliographies; nor can we deal with the programs designed to update and correct the files and to withdraw dated information.

A test of the first-generation model indicated a consistent precision or relevance ratio of over 85%—that is, of 100 items withdrawn from the shelves, less than 15 are found to be not appropriate. The explanation of this result is of course obvious. In comparing clusters of meaningful phrases found in the ABC Dictionary with information

```
                    INSTABILITY IN NEMAG* AMPLIFIER AND GENERATOR =                                ADZW
 TO REDUCE PHASE-SHIFT-DISTORTION =                    HALL-EFFECT MULTIPLIER* USING FEEDBACK AMPLIFIER AAVL
CLASS-AB PUSH-PULL AMPLIFIER =                  ANALYSIS OF HARMONIC DISTORTION* IN BALANCED AND UNBALANCED  AFIL
OR AMPLIFIER AT TWICE THE CUTOFF-FREQUENCY =          HIGH-FREQUENCY TRANSISTOR F9, PN2 AS OSCILLATOR  AAWS
                PRACTICAL 25-LBS MASER* AMPLIFIER, 20-HR HOLDING-TIME, GAIN-20, F10, F7-BANDWIDTH =       AFDB
* =                PHASE SENSITIVE 5-STAGE TRANSISTOR HOMODYNE TYPE AC AMPLIFIER AS RADIATION DETECTOR AFHX
ANCE THERMOMETER* =                                   HYBRID DC AMPLIFIER FOR THERMOCOUPLE* AND RESIST AFFB
RANSISTORS =                          DESIGN OF HIGH INPUT IMPEDANCE AMPLIFIER USING FET* AND NPN BIPOLAR T AGDM
LASH-LAMPS T300 =                       EXCITATION OF 8 INCH RUBY LASER* AMPLIFIER USING ELLIPTICAL XE F ADWX
OLAR TRANSISTORS =                       DESIGN OF HIGH INPUT IMPEDANCE AMPLIFIER USING FET* AND NPN BIP AGDM
                                                      INSTABILITY IN NEMAG* AMPLIFIER AND GENERATOR =  ADZW
S OF AMPLIFICATION AND GAIN OF FORWARD AND REFLECTED WAVE LASER* AMPLIFIER =          THEORETICAL ANALYSI AEUG
 T300 =                        EXCITATION OF 8 INCH RUBY LASER* AMPLIFIER USING ELLIPTICAL XE FLASH-LAMPS ADWX
AMIC-RANGE OF 80-DB =                                 LIMITER* AMPLIFIER OPERATING AT F7X.8 WITH A DYN AGFW
              TRANSISTOR FERRITE-CORE AMPLIFIER AS BASIC LOGIC-CIRCUIT FOR SWITCHING EQUIPMENT* =          AFWE
                                                  RUBY MASER* AMPLIFIER GAIN-14, F7-BANDWIDTH =          AEAT
10, F7-BANDWIDTH =                          PRACTICAL 25-LBS MASER* AMPLIFIER, 20-HR HOLDING-TIME, GAIN-20, F AFDB
LENT-CIRCUIT FOR AMPLIFIER DESIGN =                   MEASUREMENT OF TRANSISTOR* PARAMETERS AND EQUIVA ADLV
                                LOW NOISE TUNNEL-DIODE* MICROWAVE AMPLIFIER, F10X.1, NF5 =              AFXM
ARRIER AMPLIFIER SYSTEM =                             MINIATURIZED ELECTROCARDIOGRAM* TRANSISTORIZED C AFVM
FORMANCE OF PARAMETRIC-DIODES* IN AMPLIFIER AND FREQUENCY MULTIPLIER CIRCUITS =                    PER AFOJ
HASE-SHIFT-DISTORTION =                        HALL-EFFECT MULTIPLIER* USING FEEDBACK AMPLIFIER TO REDUCE P AAVL
-CIRCUITS USING TUNNEL-DIODES* =                      NEGATIVE-RESISTANCE AMPLIFIER AND BISTABLE PULSE AFCH
ING TUNNEL-DIODES=                DESIGN AND ANALYSIS OF F8 NEGATIVE-RESISTANCE* AMPLIFIER AND OSCILLATOR JS ABHF
                                        INSTABILITY IN NEMAG* AMPLIFIER AND GENERATOR =               ADZW
```

Figure 1. Sample from the first-generation ABC Dictionary.

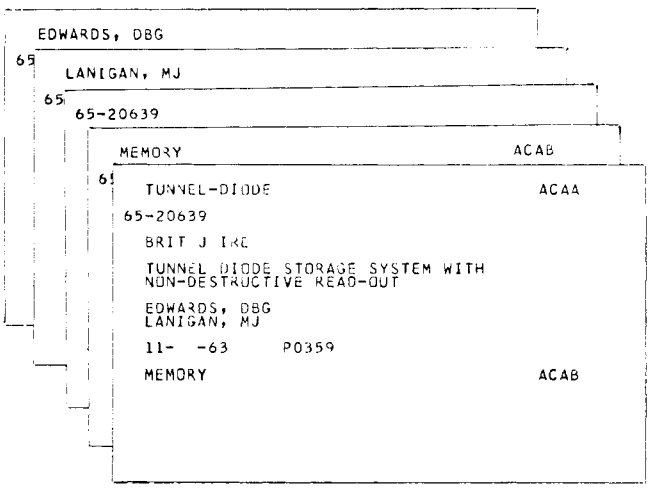

Figure 2. Complete set of catalog cards.

requirements, the searcher can pick the appropriate ones. The mental process of matching concepts of publications with those of the problem at hand is a task every scientist must perform whichever retrieval system he is using. With the ABC system, however, he compares his requirements with concepts that knowledgeable evaluators have selected for the description of publications; and the identification materials at the very beginning of the retrieval process will save time and effort of investigators as well as of personnel in the information office.

At the present time, we are designing the second-generation model. Its new features are the results of our test, as well as of a careful analysis of the operations and procedures of the system. We aim at the following objectives (2): (1) to increase the present capability of permutation from a limitation of 105 characters per sentence to one of at least 600 characters; and (2) to prepare and issue better organized ABC Dictionaries. The strictly alphabetical arrangement to the permuted descriptors (the current format) has proved its value because (a) the dictionary is completely cross-referenced; and (b) the automatic compilation is an extremely fast and economical process.

However, within the narrow semantic sectors the organization can be sometimes rough; for example, indentical terms with different inflectional forms singular and plural; and those with and without asterisks are arranged and printed out as separate groups. Screening also becomes a somewhat clumsy task when 20 or more pages of descriptor phrases are brought together under broad terms such as: amplifiers, guidance, infrared, semiconductors, oscillators, etc. Such large segments of the dictionary can be efficiently managed only through an appropriate, logical, or topical organization. We are, therefore, preparing subject schemes that will drastically reduce the time of search.

A	General—Theory and Design
AA	General Theory
AB	Threshold Conditions
AC	Spectral Lines and Spatial Distribution of Output
AD	Energy Levels
AE	Absorption and Emission
AF	Lifetime and Transition Probabilities
AG	Noise
AH	Design, Research, and Development
AI	Damage Effects Due to Lasers
AJ	Damage Effects on Lasers
AK	Effects of Doping on Basic Materials

B	Steady-State and Transient Operation
BA	Wave Analysis
BB	Mode Analysis
BC	Model Coupling
BD	Relaxation
BE	Pulsation
BF	CW Operation
BG	Output Power
BH	Efficiency

C	Resonators and Amplifiers
CA	Fabry-Perot Resonator
CB	Resonators with Spherical Reflectors
CC	Diffraction Loss
CD	Optical Waveguides
CE	Other Specific Configurations of Resonators
CF	Quantum-Mechanical Amplifier
CG	Parametric Amplifier
CH	Traveling-Wave Amplifier
CI	Other Specific Configurations of Amplifiers
CJ	Oscillators
CK	General Resonators and Amps

Instead of having to scan each line of 20 pages filled, for example, with laser descriptors, the user will concentrate his search only upon those listed under the few subtitles in which he is actually interested. Each subject scheme allows for approximately 700 subdivisions which are encoded for automatic grouping and if necessary for multiple printout of the individual descriptor sentences.

Certain descriptive terms such as Analytical Study, Bibliography, Proceedings, Design are being withdrawn from the texts of the descriptor-sentences and, by way of encoding, transferred to the entries in the subject card or book catalog. The scientist coming from the ABC Dictionary to the card catalog will therefore find the titles for a given descriptor in a definite order by type or form of publication, work phase, degree of difficulty, and selected parameters. These filter codes will contribute to a further decrease in the retrieval of unwanted publications and to a further reduction in the size of the ABC Dictionaries:

A. Analytical studies (mathematical analysis)
B. Bibliographies
C. Collection. Proceedings
D. Development, design and engineering studies and reports
E. Evaluations and tests
F. Feasibility studies
G. General, popular presentations
H. Historical studies

We will improve standard operating procedures for the preparation of the descriptor sentences, and test them within the organization and outside.

Experiments will also be conducted to obtain, through cooperation of the authors and editors in other installations, the information that can be incorporated into our system without major changes.

We will conduct studies of (a) how to produce by automatic process a thesaurus with definitions from our dictionary which, as we have seen, is a compilation of keywords in a variety of contexts; and (b) how to design an economical, mechanized method for the standardization of terminology and syntax.

We will prepare programs to retrieve documents by computer operations—that is, on the basis of three to five series ("or" and "and" combinations) of terms—without requiring the searcher to inspect the dictionary and the card file. For the sake of objectivity we must point out the two major drawbacks of the ABC system.

It is limited in the presentation of the descriptor sentences (especially in the dictionary) to the letters and symbols of the computer printer in a given installation. For the subject matter covered by our installation, this does not pose a major problem. However, information offices that wish to include chemical formulas in their ABC-type dictionaries cannot completely rely upon the standard equipment.

The cost for additional input and output devices for conventional molecular or empirical formulas seems to be relatively small. The inclusion of structural formulas into the dictionary presents greater difficulties and requires more complex and expensive procedures and equipment. Present developments and programs permit us to predict that the simultaneous retrieval of both chemical textual information and chemical formulas through an ABC-type system will be entirely feasible.

The second limitation is the more important one and requires special emphasis. Our system requires a number of subject specialists capable of evaluating and analyzing the input with respect to its particular missions and operations. At present, we have solved this problem by supplementing our in-house capability with the services of consultants, in particular, of university professors. In the future we hope through cooperation of the information-generating installations to receive descriptor-phrases and sentences in a form requiring only minor adjustments so that our efforts and expenditures can be further reduced.

Concerning the information requirements of scientists and engineers, the deductions and assumptions of our relatively cursory surveys in 1960 were corroborated by the results of a recent Department of Defense-financed investigation. It disclosed that, when in need of information, only 1% of about 1400 scientists and engineers interviewed consulted a librarian, and only 4% went to a library. Efforts exerted to locate technical data were greater than those exerted to acquire information on the state-of-the-art from reports and periodical literature. The authors of this report summed up their findings with the statement that "even greatly improved formal information centers will meet with less than a full measure of success until the user actually becomes an integral part of the system" (3).

Furthermore, our more or less intuitive evaluation of the coordinate indexing systems mentioned above was confimed by several tests, especially by the test results of the Mitre Corp.: J. F. Rial in the final report published in 1964 arrived at the conclusion that "a generally acceptable solution to the document retrieval problem must be achieved by either vastly extending the power of the basic coordinate indexing process; or by replacing this process by an altogether different one." We believe that the ABC system offers such a different process and a new approach to a hopefully satisfactory solution for scientists and for engineers.

LITERATURE CITED

(1) Altmann, B., "The Medium-Sized Information Service: Its Automation for Retrieval," TR-1192, Dec. 30, 1963 (AD 429 242).

(2) Altmann, B., "A Multiple Testing of the ABC Method and the Development of a Second-Generation Model," P. I, TR-1295, April 1965, p. 26 (AD 617 118); P. II, TR-1296, Oct. 1965 (AD 625 924).

(3) *Ibid.*, Pt. II, TR-1296, Oct. 1965, p. 19 (AD 625 924).