The ARCS System: Ringdoc as Used with a Computer*

HAROLD OATFIELD
Technical Information Services, Pfizer Medical Research Laboratories, Groton, Connecticut

Received October 20, 1966

A procedure developed for internal supplementation of Ringdoc encoded abstracts in a tape library with additional published material is described. The Automated Ring Code Search system uses a source language, ARCSLANG, for computer retrieval operations. Printouts so produced may contain codeless scanning elements as well as the usual author-title-citation. Items of specified nature may be flagged for quick identification or to form separate bibliographies generated simultaneously by the computer.

When a decision had been made in 1964 to enter a one and a half year trial subscription to Derwent's Ringdoc service, it became incumbent upon us to plan for its fruitful exploitation. Part of the reasoning behind that decision was an urge to provide our company organization with a computerized basis for retrospective literature searching. This searching capacity was desired not only to serve the obvious and expanding needs of our research staff but also those demands recently made legally binding upon us to satisfy the Food and Drug Administration requirements for appropriate bibliographies on (1) our own medicinal products and (2) related drugs. This aspect of our adaptations of Ringdoc usage is discussed here.

Out of the work to achieve Univac III printouts came the Automated Ring Code Search or ARCS system and a machine language, using Boolean "and," "or," and "but not" relations. It is a so-called source language, which means that it permits direct access to the computer without an intermediate compiler deck of instructions. The programmer, J. M. Detmer, called the process ARCTRAN and its specialized language ARCSLANG.

Figure 1 shows the position of the ARCS system and file in the Pfizer scene.

1. Sources of documents from seven Pfizer areas and a large group of non-Pfizer origin.
2. Primary documents so obtained include all types of internal periodic and special reports, external reports, and the published literature.
3. Recipients of these Documents are:
   b. Management
   b. Clinical research
   c. Research scientists
   d. Technical information
   e. ARCS System
   f. Data control
4. The Files so generated are:
   a. General central
   b. Preclinical
   c. Clinical
   d. Library
   e. ARCS File
5. Secondary documents: Fifteen types are derived from these batteries for specific purposes, which serve
6. Primary users in 11 specialized fields to generate
7. Tertiary documents that pass through stages of evaluation, commentary, management review, editing, and selection; finally, a certain portion of them is transmitted by liaison staff in nuggets to the FDA.

The Ringdoc indexing method involves encoding of chemical structures by fragmentation into units and encoding of biological concepts, including diseases, organisms, organs, tissues, etc. References coded in this manner can be retrieved by applying a search strategy, determining a search logic, and writing a search program.

Ringdoc procedure makes heavy use of "pattern cards" to simplify the step of repeated encodings for the 2000 ± most commonly cited biologically active chemical compounds. As soon as Derwent provided us with a list of these compounds together with the corresponding deck of punched cards (which was not too long after we learned to use the code), we assigned numbers to each pattern and began the compilation of "pattern books," made up of large form sheets on which were entered: (a) the name and (b) number of the compound with (c) its major trade names, (d) its structure, and (e) the specific Ringdoc coding for it (Figure 2).

Inasmuch as there are currently three distinct Ringdoc chemical codes—namely, General, Steroid, and Peptide—that compilation entailed using three separate forms. After identification of these products (not always as simple as it sounds) we eliminated duplications among them as rapidly as they became apparent to us, in order to have only one number and pattern for each chemical entity involved. This step required matching of many cards, and in some cases disclosed to us coding discrepancies.

* Presented before the Division of Chemical Literature, Symposium on User Evaluation of Secondary Sources of Chemical Information, 152nd National Meeting of the American Chemical Society, New York, N. Y., Sept. 14, 1966.

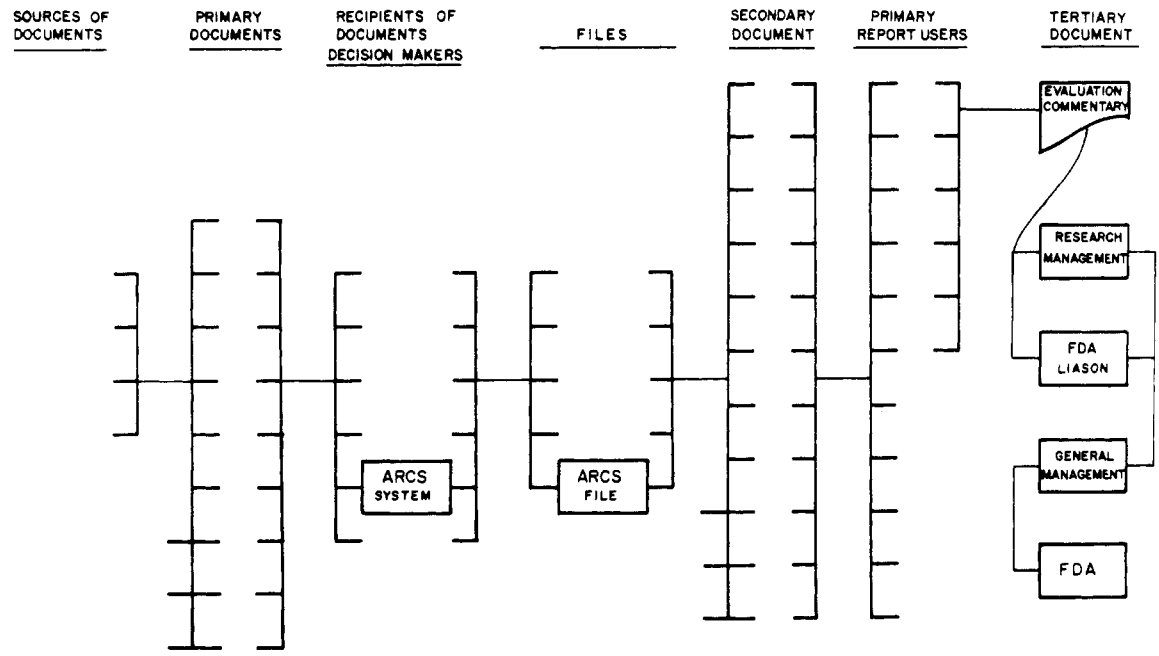

Figure 1. General flow sheet for research information.

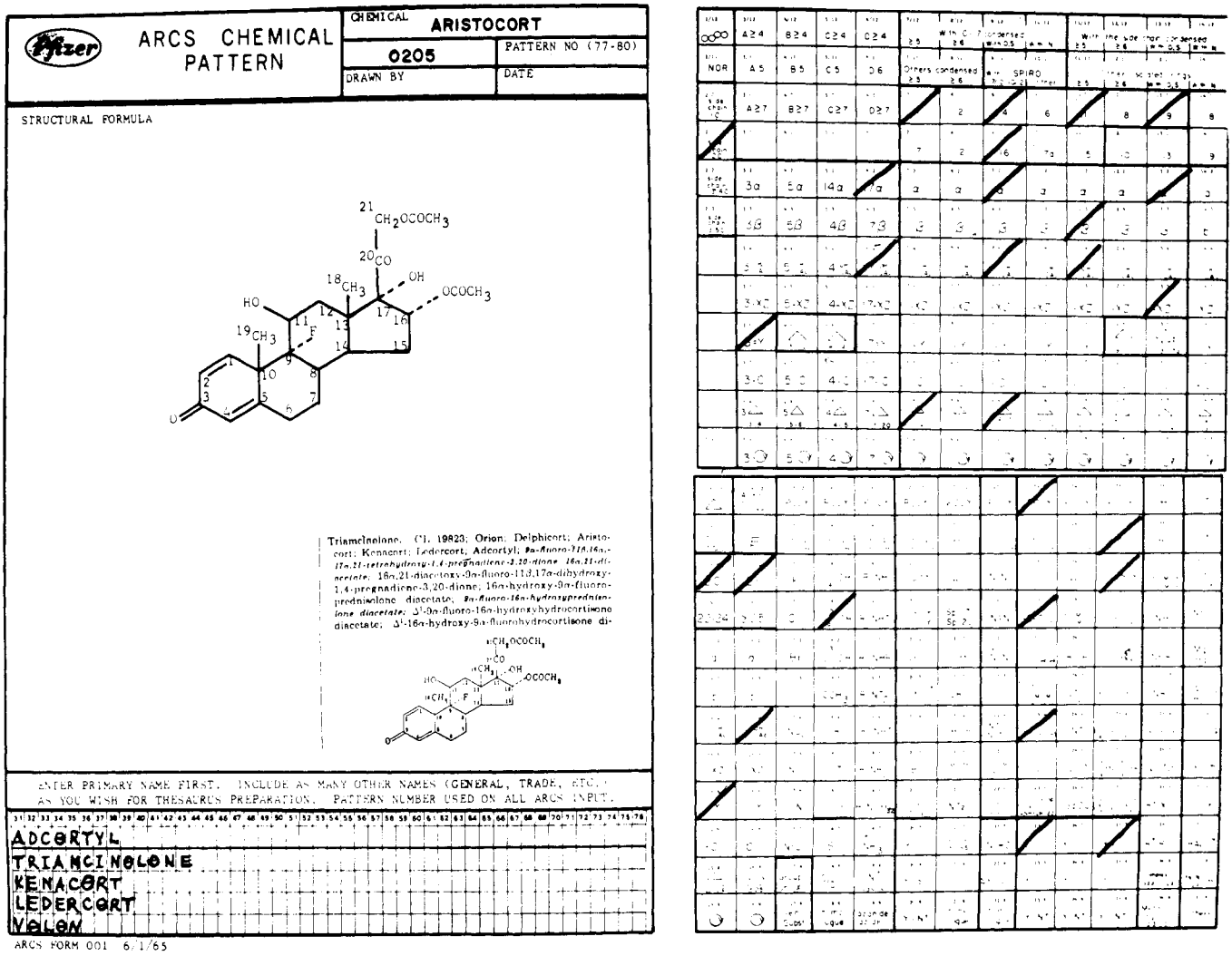

Figure 2. Pattern sheet.

From that effort, a thesaurus of compound, generic, trade, and trivial names was developed, and printed out in two forms, alphabetical and numerical by pattern number. New editions are printed at bimonthly intervals to keep it up to date. Then, to those of our products not already included in this list, we assigned individual numbers and treated them in the same fashion as new patterns. Duplicate decks of pattern cards, encompassing both Derwent and Pfizer patterns, are kept in Groton for reference, and also in the New York Data Processing Department for use in retrieval procedures.

Next, we attempted to find those drugs related to our products, within the FDA regulation's definition (as interpreted for us by our Legal Department) and make both specific patterns for these drugs and, subsequently, general patterns encompassing each series of products—*e.g.*, one for thiazides to use in connection with our Renese-*Polythiazide* reporting.

It was obvious that while Derwent's basic 237 journals held the cream of the medicinal literature, it did not hold all of it, nor even all of that portion which regularly reaches our staff people. So, we made provision for abstracting pertinent articles from about 300 additional journals received in the Pfizer libraries, and having that material encoded by the same Ringdoc code, in order to expand the published data retrievable into these special bibliographies for the FDA. We consider that the volume is now sufficient to justify reliance upon the content of the ARCS system for that purpose. About 95% of that input comes from Derwent at present.

Our abstracting procedure differs in at least four ways from that of Derwent. First, we are selective rather than all-inclusive in processing the contents of journals designated; and second, we have added to the scope of selection certain fields of company interest which are either ignored or scanted by Derwent—*e.g.*, cosmetics, flavoring adjuvants, baby formulas. The third way in which our abstracting differs from Ringdoc is in the quality of abstracting—we are much more sketchy about it. Fourth, we rely upon certain secondary sources for abstracts of additional material. Thus, we draw upon DeHaen services (*Drugs in Prospect, Drugs in Use,* and *Drugs in Combination*), *Index Chemicus, CLUE,* and *Birth Defects* for further material to incorporate. Pfizer abstracts are differentiated from Derwent's through a distinct series of accession numbers. The Data Processing Department in New York cuts cards appropriate to these internal abstracts and feeds them onto the tapes at weekly intervals.

We have changed the Ringdoc coding procedure in only one major respect, that on veterinary medical products, drastically recoding items in this category received from Derwent before introducing them onto the tape.

We have devised other working forms to use in our coding and decoding operations, but this is purely for the convenience of our electronic data processing people, and does not affect use of the Ringdoc code itself.

The over-all ARCS operation entails some 40 steps, 24 of which are clerical in nature, involving maintenance of 12 separate files and use of appropriate forms, in order to have the whole procedure under control (Figure 3). Note in particular that all the data collects or culminates in the ARCS II Master Tape File.

Figure 4 shows two groups of basic subfiles—namely, those for the abstracts themselves, and for the pattern card routine.

Figure 5 shows the component steps in the abstracting/coding procedure: running through selection of papers; preparation and editing of abstracts; concomitant bio- and chemical coding of the abstracts with verification, using corresponding forms and necessitating considerable record-keeping and filing. The abstracts generated internally will be reduced to microfilm on a yearly basis, just as the Derwent abstracts are.

Ring code punch cards on receipt from Derwent are checked in, sorted, verified for certain basic points—*e.g.*, that all alphabetic or so-called clear text terms possess a necessary 1/7 punch and read onto tape, but without a real proveout in editing. Additions, insertions, and corrections are made to the tape as necessary, and this editing causes some generation of new tapes. When possible, the codeless scanning cards and cards arising from Pfizer abstracts are batched with the Ring code cards.

An ARCSLANG program is divided into five divisions: Heading, Input, Procedure, Output, and End. The Heading explains the search in the form of remarks; the Input division employs options to reduce the data set input to the search; the Procedure contains all the logical testing; the Output employs options for selecting the type of lists or cards wanted on each match; the End signals the final statement in the Program.

Each statement contains a three position operation field (OP.), which specifies what action the computer is to take. In many statements, the OP. alone defines the action; in others a variable is required, punched in the content field (CONTENT). Furthermore, any test element may be negated by entering a minus, or dash, in the negation field (–). Finally, a tag is associated with every logical expression by making an entry in the tag field (TAG).

Figure 6 emphasizes the query procedure whereby a question submitted is restated verbally, then in ARCSLANG search logic to get computer response. This decoding step often requires exploratory diagramming to clarify the sequence of machine stages required before search cards can be cut. Among the special instructions we use are: (a) Δ to indicate that no response in a designated column qualifies, and $-\Delta$ which indicates the opposite—all responses in a column are valid; and (b) XP5 and XP7 which free us from the tight restrictions of adhering to an entire pattern, particularly with carboxylic acid derivatives and with radioactive atoms or metals in the molecule. Our questions are batched on a weekly schedule. We have been allotted up to two hours of computer time a week, and so far that amount has sufficed.

Figure 7 shows specific applications of the output from this same ARCS II master tape file. An option to have punched cards regenerated in addition to printout bibliographic citation and/or Codeless Scanning heading for the computer-selected abstracts has proved helpful for making reruns to get finer breakdown of search results, in determining reasons for false drops, and the like. At present our use of ARCS II searching runs about 60% for FDA purposes and 40% to support research efforts.

We are in the throes of adapting several aspects of our internal data reporting and control to procedures which

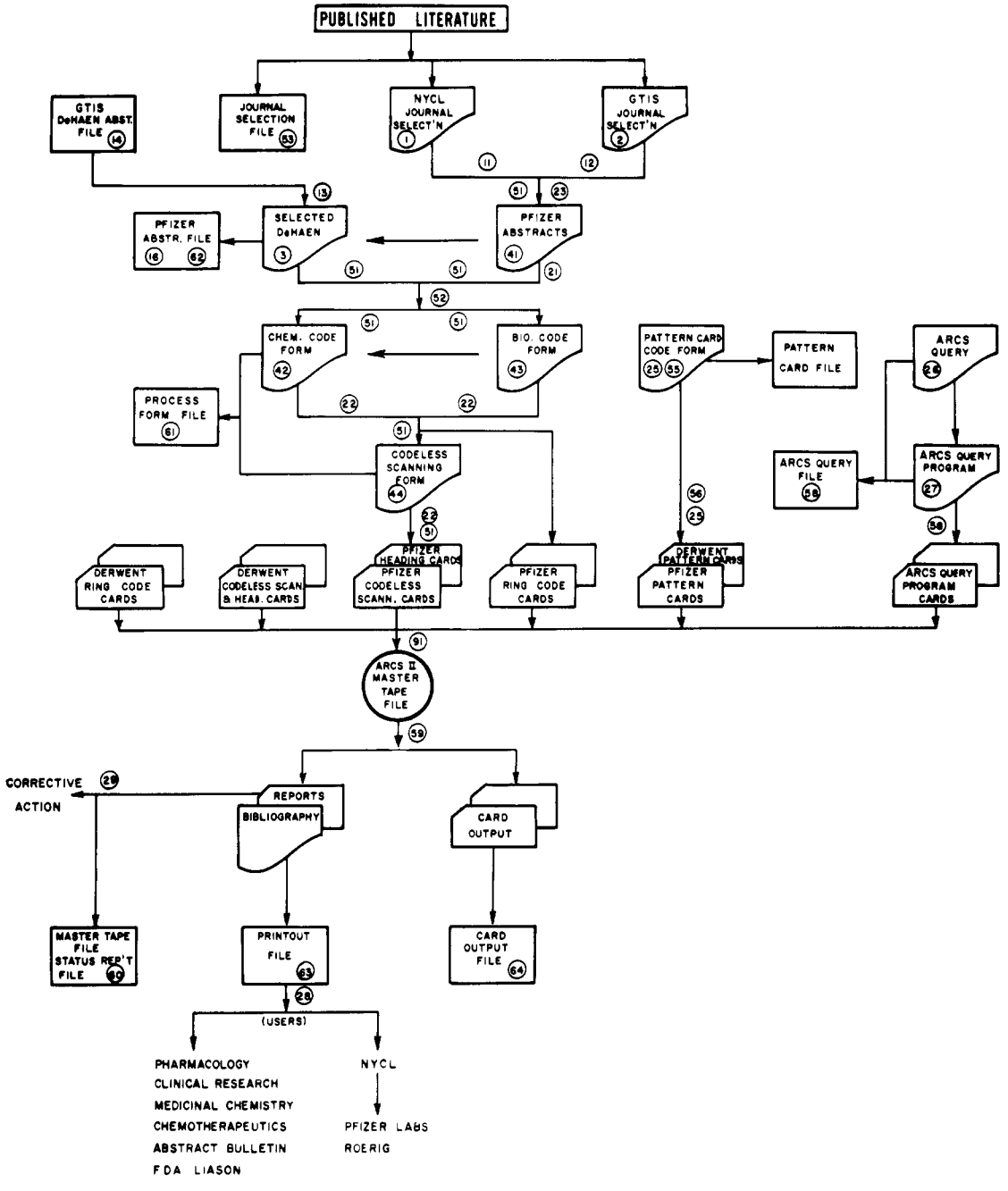

Figure 3. Flow sheet for ARCS operation in the Groton Medical Research Laboratories.

will make practicable retrieval of suitable segments by the same ARCS II Computer Program when applied to another tape library.

We draw upon both major parts of Derwent's Ringdoc service for our master tape, using the Ringdoc punched cards to provide access to the drugs and chemicals described and their pharmaco-medical applications, and the Codeless Scanning punched cards as input for the bibliographic citation.

The ARCS II program permits us to check regularly upon total content of the tapes, giving the status of specific abstracts submitted by either Pfizer or Derwent: 1. whether present or absent; 2. abstract encoded but no citation; 3. citation present but no abstract. This data control

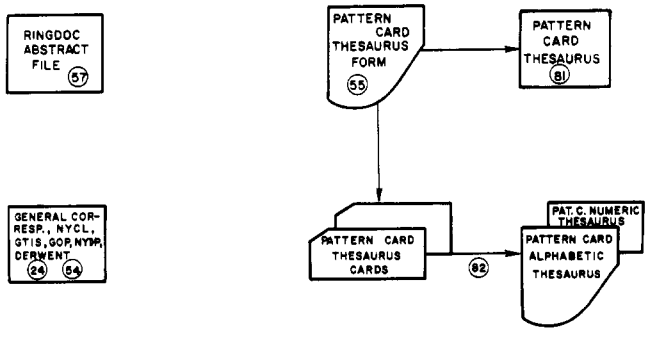

Figure 4. Two groups of subfiles: *left*, abstracts; and *right*, pattern card routine.

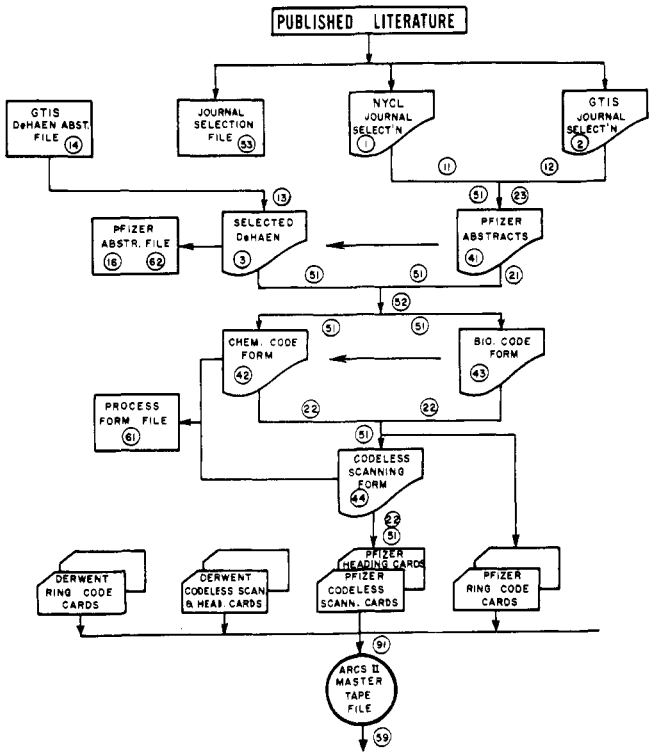

Figure 5. Abstracting coding procedure.

sheet also provides clues to errors that may have occurred, indicating their nature, within 61 categories. By a subprogram, a printout of clear text terms can be obtained, and a count made on use of each punch position for coding over a given span of abstracts recorded. The thesaurus thus generated of clear text terms that have been entered enables us to see what has been used and strive for uniformity in presentation both in spelling and spacing, and perhaps to select further appropriate terms when framing a question. The punch position distribution count helps in estimating the probable volume of responses to a query. It also guides us in forming search strategy to try to eliminate failures at the earliest moment, to streamline the search, and to reduce machine time used by putting the most restrictive requirement first. Another subprogram not yet in full working order will permit the preparation of an author index to the published literature contained in the ARCS system, and also provide breakdowns by source (institution and place of origin).

The machine printout of ARCS II contains:

1. Author(s) (up to 5 inclusive).
2. Article Title.
3. Journal Citation.
4. Codeless Scanning terms (optional). This feature was introduced into the program for the system at a late hour, after hearing from Dr. Paul Craig how valuable the feature had proved to be in the Smith, Kline and French program. We heartily agree. Its virtue arises from the fact that the

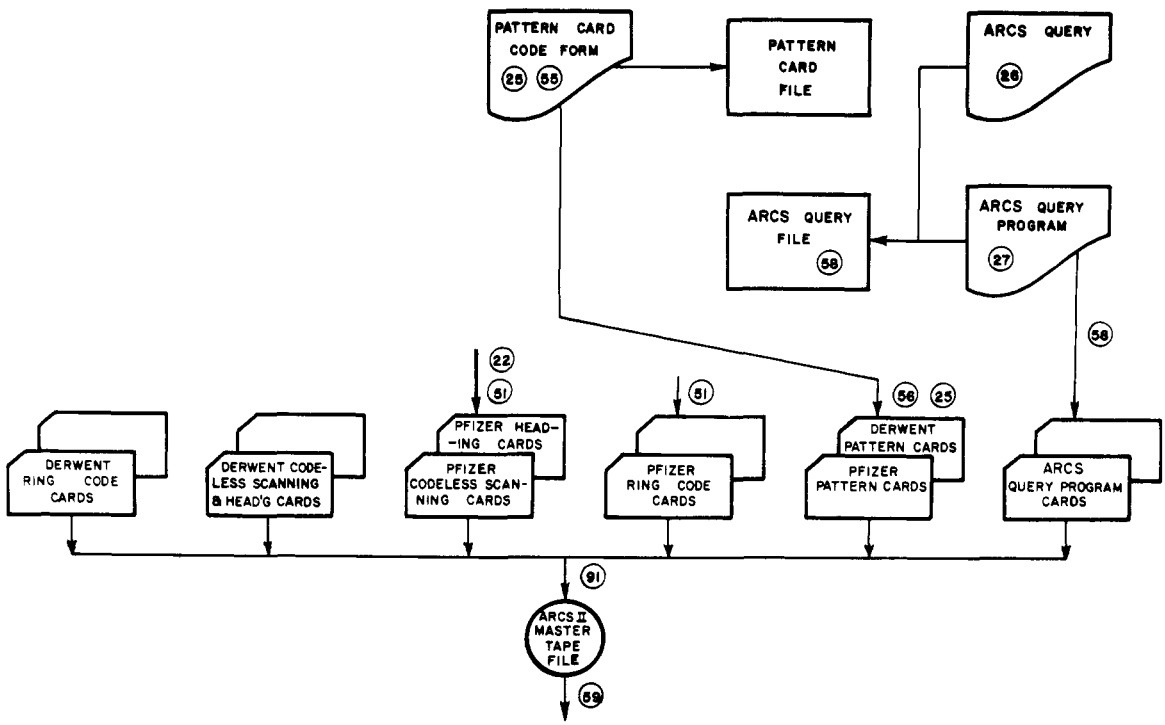

Figure 6. Query procedure.

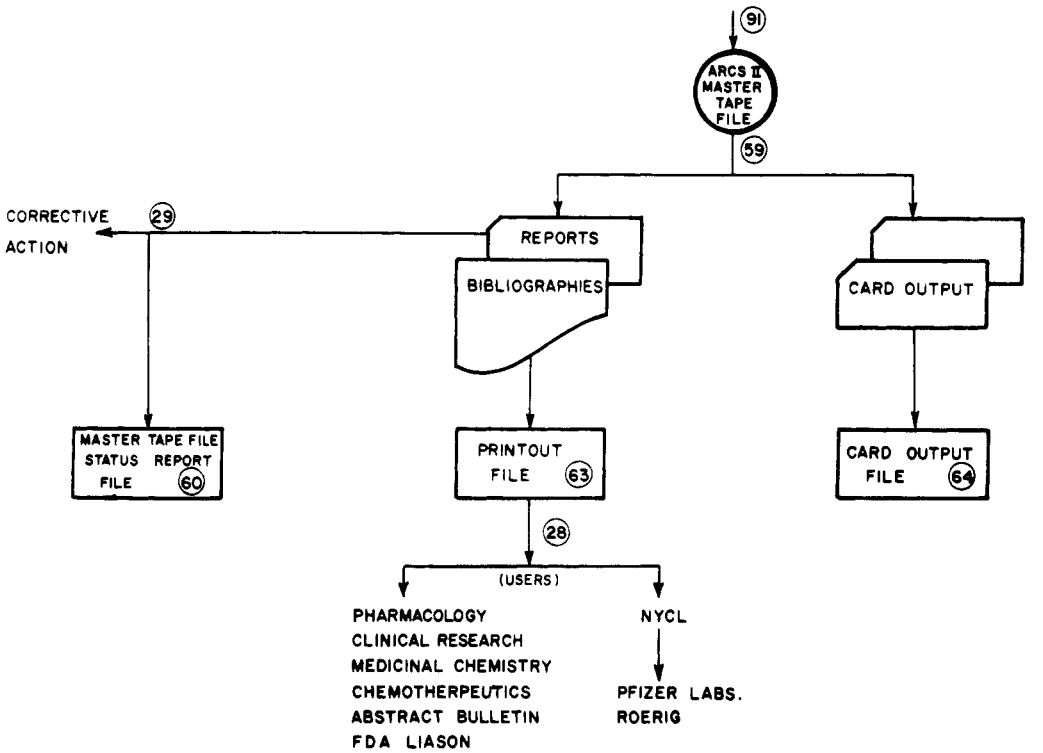

Figure 7. Output from ARCS II master tape file.

relevancy of articles machine retrieved to the query put can often be determined by scanning, as well as whether or not the abstract or the original paper should be consulted.

5. Abstract Accession Number. (The ARCS I Program, developed as an interim procedure retrieved and printed out only abstract numbers).

6. Total number of references (on first page, together with reference data on the search cycle, query number, requester, the question expressed in both English and machine language, limitations, if any in scope†, decoder's initials, and a key to flags used). We have at our disposal eight possible flags, or indicators, for selected types of data, chemicals, or concepts. Three of these flags are always pre-empted in any FDA searches by markers respectively for Pharmacology, Therapy, and Teratology. Whenever an FDA output instruction has been entered for a query, a unique procedure is automatically triggered when the last ring code for each abstract number is read. At this signal the ring code cards which have been stored in core are scanned and a flag is set for therapy if there are any punches in columns 72 or 74 or mark the presence of 5, 6, or 7 in column 73. A flag is also set for pharmacology if there is a punch in any of the first seven positions in column 73. An FDA output instruction will result in separate printouts of pharmacology and therapy bibliographies from a given search. The printout may be used as it stands, or the requester may indicate items he wishes deleted from it and submit if for machine reprinting. The order of items cannot be changed, except through a new search with resultant new printout.

Pfizer veterinary coding procedure differs in two respects from Derwent's:

1. The range of animals treated in veterinary medicine is much wider than that encompassed in pharmacological procedures. Therefore, whenever we encounter the veterinary medicine punch, the significance of card columns 36 and 37 is arbitrarily changed. It designates a different series of creatures in the two columns devoted to test object.

2. Derwent applies the WHO disease code (*1*) not only to therapy of human beings but to treatment of animals, insofar as analogous terms apply. Pfizer does not regard that procedure as valid. Hence, we use the U.S. Dept. of HEW's newly revised *Standard Nomenclature of Veterinary Diseases and Operations* (*2*). This code uses a block of nine successive numbers or letters. To muster sufficient card columns for that purpose, we have taken the unused columns 63 and 64, the alpha-numeric columns 65 and 66, and the five succeeding columns normally reserved for the disease code. Because the only letters involved in the first eight postions are X and Y, this coding can be entered directly by using the zone punch positions and the nine numbers. Although the fine coding in this nomenclature table given by the ninth position, which employs either numbers or more letters, probably offers greater refinement in recording data than we will require, it is still possible to enter it in the column by conventional methods. Whenever the veterinary article also reports data which are normally entered in alpha-numeric form in columns 65 and 66 (galenicals, controls, and the like) a second card appropriately punched for the abstract records that fact.

† An arbitrary limit of 1000 citations to be printed out per query has been set in order to avoid wasteful activity when queries may have gone astray, or a search has unexpectedly proved over-fruitful. The limit can be removed when necessary, or further reduced.

Searching restrictions may also be set in the instructions at the beginning of each query in two ways. The first method is to designate an initial abstract number at which the search shall start, or a range of abstract numbers within which the search is to be conducted:

ABS00005E-99999F

The second method is to set a cycle limitation or data control:

CYC 033-060

Each weekly searching of the ARCS tapes has a cycle designation. In these ways it is possible to rerun a query in order to pick up references introduced after a previous search without extensive duplication of entries in the printouts.

At this point we had been learning how to retrieve information from this bank of literature for 13 months. We have used it as the principal means of satisfying FDA requirements for bibliographies on our drugs. To date we have submitted about 300 questions to the computer. For the FDA application the results seem to be entirely satisfactory. On research questions, aside from such conspicuous limitations as the short time span of publications involved and restricted journal base, we have found others. For instance, the code does not permit one to distinguish between α- and β-adrenergic blockers.

In evaluating results with our non-FDA retrieval efforts, there appear to be about 15% failures. A further 15% gave only fair results, in that part of the material retrieved was highly pertinent, but some other valid material known to be present was not recovered.

Among the failures were such topics as

(a) factors affecting blood flow in adipose tissue;
(b) studies concerned with the electrical charges on or electrophoresis of the formed elements of the blood;
(c) tranquilizers used in preoperative therapy of asthmatic patients.

We suspect it was lack of data rather than faulty search strategy which accounted for no response in this last case. However, a "no response" result in many of those searches was quite an appropriate reply, and often comforting.

A primary advantage of the Ringdoc-ARCS system is its timeliness in making published information available to users, and especially its machine retrievability in different contexts. The timetable which we and Derwent strive to maintain is:

Abstracts issued four weeks after original journal publication. Ringdoc cards shipped at six weeks via air freight, and introduced into ARCS tape by seven weeks, so that the material selected becomes retrievable at the eighth week following publication.

In actual practice, there have been delays of one sort or another, which have prevented achievement of that schedule entirely (shipping strikes, loss of material, faulty processing). But it is run by human beings for other no less fallible human beings with not all conditions subject to exact control.

Of course, our own present lack of full grasp of the system's capacities, and the machine's intransigence toward inaccurate, incomplete, or otherwise inadequate instructions, has pulled us up short at embarrassingly frequent intervals, too. We do feel that we are learning, that we can live with this system, refine it, and continue to perfect our use of it.

LITERATURE CITED

(1) "Manual of the International Statistical Classification of Diseases, Injuries or Causes of Death," 1955 revision; World Health Organization, Geneva, 1959.
(2) Epizootiology Section, National Cancer Institute, Bethesda, Md., 1966; Available from Superintendent of Documents, Government Printing Office, Washington, D. C., $3.50.

A Chemically Oriented Information Storage and Retrieval System. I. Storage and Verification of Structural Information

CARLOS M. BOWMAN, FRANC A. LANDEE, and MARY H. RESLOCK
Computation Research Laboratory, The Dow Chemical Company, Midland, Michigan

Received December 29, 1966

A computer-based system has been designed to handle chemically oriented files. The Wiswesser line-formula chemical notation is a practical method for representing structural formulas for input. The file organization and methods of verifying the accuracy of the notation as well as cost are described.

The problem of organizing and indexing chemically oriented data and information is a difficult one. The literature itself is quite voluminous on the subject. In 1964 an excellent survey of the various methods was published (*1*, *2*). Most of the systems described reported their result in the retrieval of the names, structures, or identification numbers of compounds which satisfy certain structural relationships. A computer-based system has been devised which will allow searches to be made not only on structural considerations but also on properties and other pertinent information about chemical compounds (*3*). The system will be described in a series of papers, of which this is the first.

This paper will discuss the establishment of the section of the file which contains the structural considerations—*i.e.*, name, structural configuration, molecular formula, etc. It will also discuss a computer program which has been written to check the accuracy of structural and molecular formula information. Finally, some typical costs of input will be presented.