ED 026 244 SE 004 966

Annual Report of the National Science Foundation on Contract NSF-C414 Task III July 1966 through June 1967.

American Chemical Society, Columbus, Ohio. Chemical Abstracts Service.

Pub Date Jun 67

Note-56p.

EDRS Price MF-$0.25 HC-$2.90

Descriptors-*Abstracts, *Annual Reports, *Chemistry, Information Dissemination, Information Processing, *Information Science

Identifiers-American Chemical Society, Chemical Abstracts Service, Columbus, National Science Foundation, Ohio

This Annual Report describes in detail the work performed during the first year of Task III of Contract NSF-C414 and the present status of Task III work. The programs and achievements described constitute the first significant efforts to develop a user-oriented, cooperative program between major secondary scientific and technical information services--the Chemical Abstracts Service (CAS) information system and the National Library of Medicine's (NLM) MEDLARS--in conjunction with a large user of chemical and bio-medical information, the Food and Drug Administration (FDA). Experimental and developmental efforts have resulted in three new computer systems being instituted to produce the NLM Output Tape, the Desktop Analysis Tools, and to determine and assign automatically terms for MEDLARS. In addition, CAS has performed 59,698 registrations. These have contributed data on 21,110 substances that were new to the Chemical Abstracts Service files. (BC)

# ANNUAL REPORT

# to the

# NATIONAL SCIENCE FOUNDATION

## on

## CONTRACT NSF-C414 TASK III

## July 1966 through June 1967

CHEMICAL ABSTRACTS SERVICE

AMERICAN CHEMICAL SOCIETY

Columbus, Ohio

30 June 1967

# ANNUAL REPORT

to the

NATIONAL SCIENCE FOUNDATION

on

CONTRACT NSF-C414 - TASK III

July 1966 through June 1967

CHEMICAL ABSTRACTS SERVICE

AMERICAN CHEMICAL SOCIETY

Columbus, Ohio

30 June 1967

# CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# ABSTRACT

This Annual Report describes in detail the work performed during the first year of Task III of Contract NSF-C414 and the present status of Task III work.

The programs and achievements described constitute the first significant efforts to develop a user-oriented, cooperative program between major secondary scientific and technical information services--the Chemical Abstracts Service (CAS) information system and the National Library of Medicine's (NLM) MEDLARS--in conjunction with a large user of chemical and bio-medical information, the Food and Drug Administration (FDA).

Experimental and developmental efforts have resulted in three new computer systems being instituted to produce the NLM Output Tape, the Desk-top Analysis Tools, and to determine and assign automatically MeSH Class Terms for MEDLARS. In addition, CAS has performed 59,698 registrations. These have contributed data on 21,110 substances that were new to the Chemical Abstracts Service files.

# INTERFACE AND INTERACTION

Chemical Abstracts Service (CAS) of the American Chemical Society is presently developing the Chemical Compound Registry System under Task I of Contract NSF-C414. The purpose of this computer-based system is to identify uniquely chemical compounds and to store and retrieve structural descriptions, chemical nomenclature, and significant literature citations for these compounds.

In contrast to the overall Task I objective of building a Registry System, Task III, which is supported by the National Science Foundation, The Food and Drug Administration (FDA), and the National Library of Medicine (NLM), has as its purpose the experimental operational interlinkage of the NLM MEDLARS and the CAS information systems, two of the world's largest secondary information services, and FDA, a large governmental administrative agency that routinely uses the secondary services to perform its mission.

The work described in this report constitutes the first significant effort to develop a user-oriented, cooperative program between major secondary scientific and technical information services that are devoted to closely related subjects and that overlap significantly in the source documents covered by each. The project, should it leave the experimental stage and become fully operational, offers the prospect of greatly simplified joint utilization of NLM and CAS information services by individuals and organizations which require regular access to both bio-medical and chemical information. Equally important, the successful conclusion of this project offers potential operational savings both for the cooperating secondary services and their large users.

Overall, NSF Contract C414 was established to develop a Chemical Compound Registry System. The Registry System is conceived to be a computer-based recognition and file system that provides the bridge between atomic and molecular structural characteristics and the corresponding systematic and nonsystematic nomenclature that appears across the entire range of scientific and technical literature. In addition, the system directly interlinks the data contained in it to the source docur.ents in which the information was initially reported.

The Registry System operates by assigning a unique number, called a Registry Number, to each substance when it is first entered into the file. Whenever a substance already registered appears in a new reference, the previously assigned number is automatically recovered. The System has three associated computer files--the Structure, Nomenclature, and Bibliography Files--within each, the Registry Number functions as a machine address to tie together all information related to a given substance.

Much progress has been made on Task III. During the first year, 59,698 registrations were performed.* This number represents the machine registerable substances from contract-specified sources. Still remaining are substances that must be manually registered because registration conventions have not been established as yet.

*A registration is defined as the process of determining the existence or nonexistence of a substance in the Registry File.

Also as a consequence of Task III, three new computer programs were developed. These included the routines required to produce the NLM Output Tape (including the Combination Record), specialized Desktop Analysis Tools (DAT) for FDA and NLM, and to determine and assign automatically MeSH Class Terms for MEDLARS, all of which are discussed in this report.

The work on this contract has progressed in great measure because of liaison efforts on the part of the parties involved. An example of such cooperation are the contributions made by personnel of each organization toward the technical upgrading of certain source lists to the level required to process them into the Common Data Base.

A similar example is the development of the MeSH Class Term Assignment System described in this report. This System is based on an adaptation of the CAS Substructure Search System. It is expected that the developed system will continue to play a useful part in MEDLARS operations and that the joint NLM-CAS effort in defining substructures will be useful to MEDLARS users. It is interesting to note that the MeSH terms which are equivalent to structural fragments in themselves constitute a useful fragment code which may be of interest to MEDLARS users for their proprietary files.

In addition to regular working sessions between FDA, NLM, and CAS staff, a technical training session was held to acquaint a group of FDA chemists with the CAS Chemical Compound Registry System. A description of this training session is given later in this report.

# ESTABLISHING THE COMMON DATA BASE

A substantial portion of the work performed under Task III of NSF-C414 was the registration of chemical substances from the sources listed in the contract. The structural and nomenclature information concerning these compounds and the less-than-fully defined substances which together comprise the Common Data Base are now for the first time available in one place in a computer-readable file.

## SOURCES OF COMMON DATA BASE

During the first year of Contract NSF-C414, Task III (1 July 1966 through 30 June 1967), 59,698 registrations were performed. Of these, 21,110 resulted in new substances being added to the CAS Chemical Compound Registry System, while 38,588 substances matched compounds already on file. Of the sources inspected and analyzed, most were books and journal articles, including four reference works whose contents were registered under Task I of this contract, but are included as part of the Common Data Base.

Table I gives a complete summary of the registration performed during the first year of this contract.

TABLE I

SUMMARY OF REGISTRY FILE

- 1 July 1966 through 30 June 1967

| SOURCE | | REGISTRATIONS PERFORMED | | |
|---|---|---|---|---|
| Name | Code | New to File | Matching Those on File | Total |
| Code of Federal Regulations[1] | CFR | 1041 | 3042 | 4083 |
| Colour Index[2] | CI | 3974 | 3024 | 6998 |
| Common Names for Pesticides | CNP | 4 | 87 | 91 |
| Dangerous Properties of Industrial Materials | DPIM | 25 | 74 | 99 |
| Drug and Cosmetic Catalog | DCC | 187 | 1303 | 1490 |
| Drug File (CAS internal file) | | 3079 | 6702 | 9781 |
| Farm Chemicals Handbook | FCH | 103 | 454 | 557 |
| Feed Additive Compendium | FAC | 105 | 154 | 259 |
| Food Chemicals Codex | FCC | 95 | 423 | 518 |
| Guide to Chemicals Used in Crop Protection | GCWCP | 31 | 412 | 443 |
| Handbook of Toxicology | HT | 264 | 226 | 490 |
| International Encyclopedia of Cosmetic Material Trade Names | IECMTN | 2255 | 2172 | 4427 |
| International Non-Proprietary Names | INN | 100 | 907 | 1007 |
| International Pharmacopeia | IP | 60 | 519 | 579 |
| List of Colors, Appendix | LC | 28 | 90 | 118 |

| SOURCE | | REGISTRATIONS PERFORMED | | |
|---|---|---|---|---|
| Name | Code | New to File | Matching Those on File | Total |
| Merck Index of Chemicals and Drugs(2) | Merck | 8416 | 12090 | 20506 |
| MeSH Terms | MeSH | 110 | 810 | 920 |
| Mycotoxins in Foodstuffs(1) | MFS | 63 | 82 | 145 |
| The National Formulary | NF | 164 | 672 | 836 |
| New Drugs | ND | 30 | 305 | 335 |
| The Pharmacopeia of the United States of America | USP | 71 | 498 | 569 |
| Perfumes, Cosmetics, and Soaps | PCS | 323 | 1390 | 1713 |
| Pesticide Index(2) | PI | 154 | 748 | 902 |
| "Pesticides" from Chem. Week | CW | 57 | 332 | 389 |
| South African Medical Journal | SAMJ | 9 | | 15 |
| Summaries of Pesticide Toxicity | SPI | 11 | 98 | 109 |
| United States Adopted Names(2) | USAW | 89 | 769 | 858 |
| Veterinarians' Blue Book | VBB | 262 | 1199 | 1461 |
| TOTALS | | 21110 | 38588 | 59698 |

(1) Substances registered only from selected and specified sections. See Appendix B.

(2) Registered under Task I.

## METHODS AND PROCEDURES

Since the Common Data Base (CDB) File is an integral part of the Chemical Compound Registry System, the registration process for CDB substances follows the procedures already established for the Registry System. Processing begins with a professional review of the sources from which substances are registered in order to select and code applicable nomenclature. These data are then clericallv keyboarded via the Mohawk 1181 Data Recorder and processed through the computer-based, name-matching system. This system compares the names of compounds being registered (usually author-assigned or "trivial" names) against the names of compounds already registered. Each time an exact match is achieved, the formerly registered compound's Registry Number and molecular formula are retrieved, and together with the input name are printed on Data Sheets for a chemist's review. Following any corrections made by the chemist, the data is added to the master Registry files along with the appropriate source codes.

Compounds for which there is no name match are registered by structure. Structure diagrams drawn and reviewed by a staff chemist are clerically keyboarded using the Mohawk 1101 Data Recorder or a chemical structure typewriter. Substances for which the structure is unknown and substances which are not now being automatically registered are assigned a Registry Number by a chemist. The compound's nomenclature, molecular formula, and source codes are associated with its Registry Number and input to the computer file. Figure 1 illustrates the data flow for Task III substances within the CAS Chemical Compound Registry System.

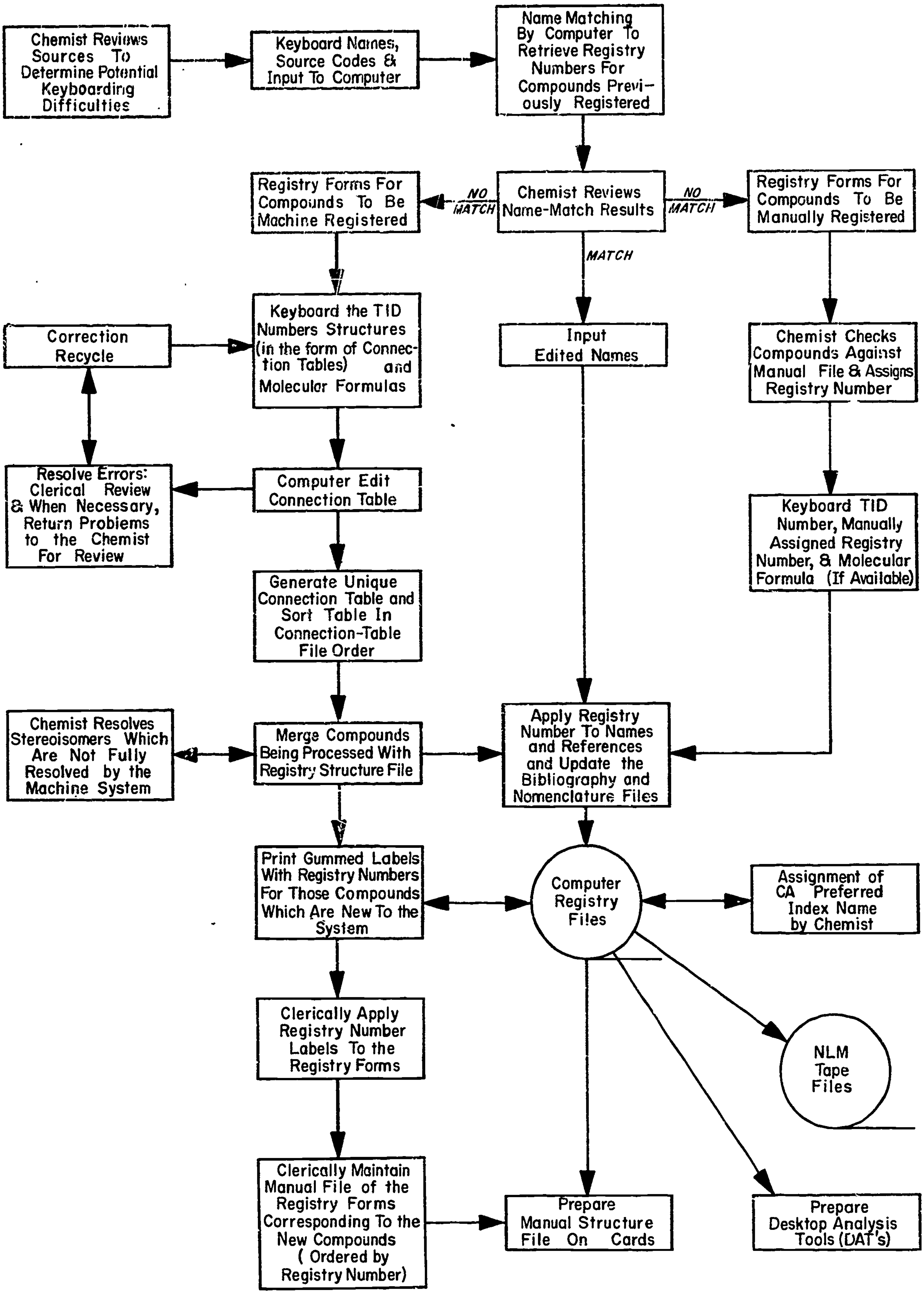

Chemist Reviews Sources To Determine Potential Keyboarding Difficulties
Keyboard Names, Source Codes & Input To Computer
Name Matching By Computer To Retrieve Registry Numbers For Compounds Previously Registered
Registry Forms For Compounds To Be Machine Registered
NO MATCH
Chemist Reviews Name-Match Results
NO MATCH
Registry Forms For Compounds To Be Manually Registered
MATCH
Correction Recycle
Keyboard the TID Numbers Structures (in the form of Connection Tables) and Molecular Formulas
Input Edited Names
Chemist Checks Compounds Against Manual File & Assigns Registry Number
Resolve Errors: Clerical Review & When Necessary, Return Problems to the Chemist For Review
Computer Edit Connection Table
Keyboard TID Number, Manually Assigned Registry Number, & Molecular Formula (If Available)
Generate Unique Connection Table and Sort Table In Connection-Table File Order
Chemist Resolves Stereoisomers Which Are Not Fully Resolved by the Machine System
Merge Compounds Being Processed With Registry Structure File
Apply Registry Number To Names and References and Update the Bibliography and Nomenclature Files
Print Gummed Labels With Registry Numbers For Those Compounds Which Are New To the System
Computer Registry Files
Assignment of CA Preferred Index Name by Chemist
NLM Tape Files
Clerically Apply Registry Number Labels To the Registry Forms
Clerically Maintain Manual File of the Registry Forms Corresponding To the New Compounds (Ordered by Registry Number)
Prepare Manual Structure File On Cards
Prepare Desktop Analysis Tools (DAT's)

*Figure 1. DATA FLOW FOR TASK III SUBSTANCES*

# INPUT DATA

For every compound registered under Task III, seven data items are recorded, if they are available. Two, the Registry Number and Item Number (see Glossary, Appendix A for definitions) are normally assigned by the computer (except for compounds manually registered for which a chemist assigns the Registry Number). Input during registration are the 1) structure, 2) molecular formula, 3) nomenclature, 4) source code(s), 5) name-type code(s). Not all these are entered for each substance, for in certain cases the data may not be known. For example, there are many natural products for which no structure has been established. Under these circumstances, the substances would be manually registered without a structure, and in certain cases, without even a molecular formula. However, all other information concerning this product would be input and tied together by the Registry Number.

The following is a brief description of each data item entered for Task III substances.

## Structures

Structural descriptions, when available, are entered into the record as atom-bond connection tables. Devices such as a Mohawk 1101* or a structure typewriter are used for this purpose. In the process that utilizes the Mohawk 1101, a clerk numbers each nonhydrogen atom in the structure diagram, then keyboards a connection table that lists each atom by number and indicates the atoms to which each atom is bonded and the type of bond involved. The computer then edits the table and automatically converts it into one that is compact, unambiguous, and unique. This latter table is compared with the master structure file and the previously assigned Registry Number is retrieved for compounds that

*The Mohawk 1101 is a keypunch that records data directly on magnetic tape.

match. A Registry Number is assigned by the computer to each compound that is new to the file. The new table is then added to the master structure file.

The second method for registering structures utilizes the structure typewriter. Here a clerk copies the chemical structure directly using the typewriter, and the computer converts the structure to a connection table identical to the one that would have been generated by a connection table entry.

Molecular Formula

The molecular formula calculated from the connection table and recorded on the computer files is the modified-Hill format used in Chemical Abstracts (CA). Computer programs have been written to convert the modified-Hill format to the form desired by NLM, an inverted "NOPS sequence" molecular formula. In the latter form, nitrogen, oxygen, phosphorus, and sulfur (if these elements are present) are listed first, followed by an alphabetical listing of the elements excluding carbon and hydrogen, which appear last. For example, glycine is represented by $NO_2C_2H_5$; the corresponding modified-Hill representation for this same compound is $C_2H_5NO_2$.

Nomenclature

The Nomenclature File of the Common Data Base (CDB) is comprised of all the names contained in the CAS files for a particular substance. For retrieval purposes they are categorized and identified in the following manner:

- CA Preferred Index Name (inverted form)
- CA Index Names (uninverted form)
- Added CA Index Names
- "Trivial Names"

Names and synonyms are selected not only from the designated sources, but also from data already in the Registry Nomenclature File. When a compound is registered, and a match found, synonyms already on record are included as part of the Common Data Base.

The four specific codes used for the names incorporated into the file for Task III substances are detailed below:

1. *CA Preferred Index Names* are the systematic names used in the Subject and Formula Indexes of *CA*. Each substance is assigned such a name according to the established *CA* nomenclature policy at the time. Although no formal plan exists to update these names once they have been assigned, through use, some limited updating has been made throughout the year.
2. An *Uninverted CA Index Name* is the uninverted form of the *CA* Preferred Name. It is the verbal form and as such is the form usually recognized by chemists. For example, "hexachlorobenzene" is the uninverted form of the *CA* Preferred Name, "benzene, hexachloro-."
3. *Added CA Index Names* are special-purpose names that emphasize special structural features and are given in addition to the *CA* Preferred Index Name for a substance in the *CA* Subject Index.
4. *"Trivial" Names* are names derived by nomenclature rules other than those used for the *CA* Preferred Index Name. In addition, "trivial" names include such designations as laboratory numbers and trademarked names.

## Registry Number

The Registry Number is a unique computer-checkable number assigned to each substance when it is first entered into the file. Whenever a substance which is already in the file is registered, the previously assigned number is recovered automatically. The Registry Number functions as a machine address within the files of the Registry System to link together all information about a given substance.

## Source Code(s)

Source codes have been assigned to all of the sources from which the substances are registered for entry into the Common Data Base. These codes are output as part of the data from the Registry System to facilitate the use of the system.

## Name-Type Code

The Name Type identifies the type of name as follows:

| Code | Type |
|---|---|
| 1 | Preferred *CA* Index Name (inverted form) |
| 2 | Added *CA* Index Name (inverted form) |
| 3 | Author Name or Trivial Name, including Laboratory Numbers |
| 4 | Preferred *CA* Index Name (uninverted form) |
| 3R | Registered Trade Mark |
| M | Unique MeSH Term |

## Item Number

The Item Number is an identification number used in conjunction with the Registry Number to provide a unique means of accessing data in the CAS Bibliography and Nomenclature Files. Each Item Number appears as a two-to-six character sequence, the last character being a check digit or letter used for automatic checking of the full series of digits obtained by combining the Registry Number and the Item Number. As an example, the name Benzoic acid, 2,4-dihydroxy- is item 3 associated with Registry Number 89861 on the CAS Nomenclature File. The check digit for this item is 6, which is used to verify the entire sequence, 89861 3. The sequence of digits recorded is thus: 89861 36.

# SYSTEMS DEVELOPMENT

As previously discussed, Task III of contract NSF-C414 is an experimental program directed toward the development of computer-based outputs that interface specifically with NLM and FDA information programs. To meet the requirements of this specialized orientation, CAS has developed and has placed into pilot plant operation the three new computer-based systems listed below.

- The NLM Output Tape System (including the Combination Record)
- MeSH Class Term Assignment
- Desktop Analysis Tool System

The original development and the continuing improvement of these programs represent a substantial investment in time and money. The first two systems are described below. The Desktop Analysis Tool System is detailed in the section entitled "Output".

## THE NLM OUTPUT TAPE SYSTEM

The National Library of Medicine (NLM) requires that data on each registered substance include available names, molecular formula, MeSH Terms, and components (if the substance is a mixture) as well as Registry Number, source codes, etc. It also requires that no data for a substance be sent to them unless all the necessary data items are available. In addition, the taped information must be suitable for processing on their Honeywell 200 computer. To meet these requirements, CAS has developed what is referred to as the "NLM Tape Output System," a computer system that detects, extracts, holds, and reformats Common Data Base data to be forwarded to NLM. The development of this system has required a great deal of communication between NLM and CAS to obtain output in the exact format required by NLM. Although the basic programs have been completed for some time, many modifications have had to be incorporated into the routines because of initial mutual misinterpretations and misunderstandings concerning formats and design specifications. However, the liaison between NLM and CAS personnel has served to clarify these needs, and the NLM Tape Output System is well on its way to meeting NLM's needs.

The total Tape Output System is comprised of two subsystems, the Pending System and the Reformat Program. The former performs essentially a monitoring function; that of holding data about a substance until all the required information concerning that substance is processed into the File and is available as a unit for release.

An illustration will serve to describe the operation of the Pending File: data for Substance X is keyboarded into the Bibliography File. The information is comprised of the Registry Number, name, and molecular formula, but no MeSH Term. This type of event frequently occurs because the various data items are input and processed independently of each other and follow separate processing routes to the Pending File. When the initial data are received by the File, the computer recognizes that not all the required information concerning Substance X is available and therefore the recorded data are held. When later the MeSH Term is added to the initial data, completing the record, the unit is released for further processing.

Material released from the Pending File is next fed into the Reformat Program which performs two functions. The first is rearrangement whereby the data such as names, item numbers, molecular formulas, and MeSH Terms are recorded in the sequence desired by NLM. Second, the computer program converts the nine-track designation of the IBM 360 format into the seven-track format of the Honeywell computer used by NLM. If, for instance, the designation for the letter "A" in the 360 language is 01100100 and the designation for the letter "A" in the Honeywell 200 format is 001010, the program automatically changes the former into the latter.

Once the NLM Reformat Program has made the two changes described, the information is duplicated onto another magnetic tape which is sent to NLM. The original tape is retained by CAS until the next output tape is produced to prevent the information from being lost in transit. The information for each substance registered contains the four records mentioned above, namely, a name record, molecular formula record, MeSH Term record, and a component record. This latter record contains component Registry Numbers if the substance is a mixture and the components are identified in the literature. If the components are not known such as in the case of certain "oils," the record is filled with zeros.

In addition to the records described above, the NLM Tape contains also all the identification data such as item number and Registry Numbers that serve to tie all the information together. A detailed flow diagram of this system is shown in Figure 2.

## The Combination Record

The Combination Record is that portion of the NLM Output Tape System that records on magnetic tape the Registry Numbers and cross references of mixtures and related components (See Figure 2).

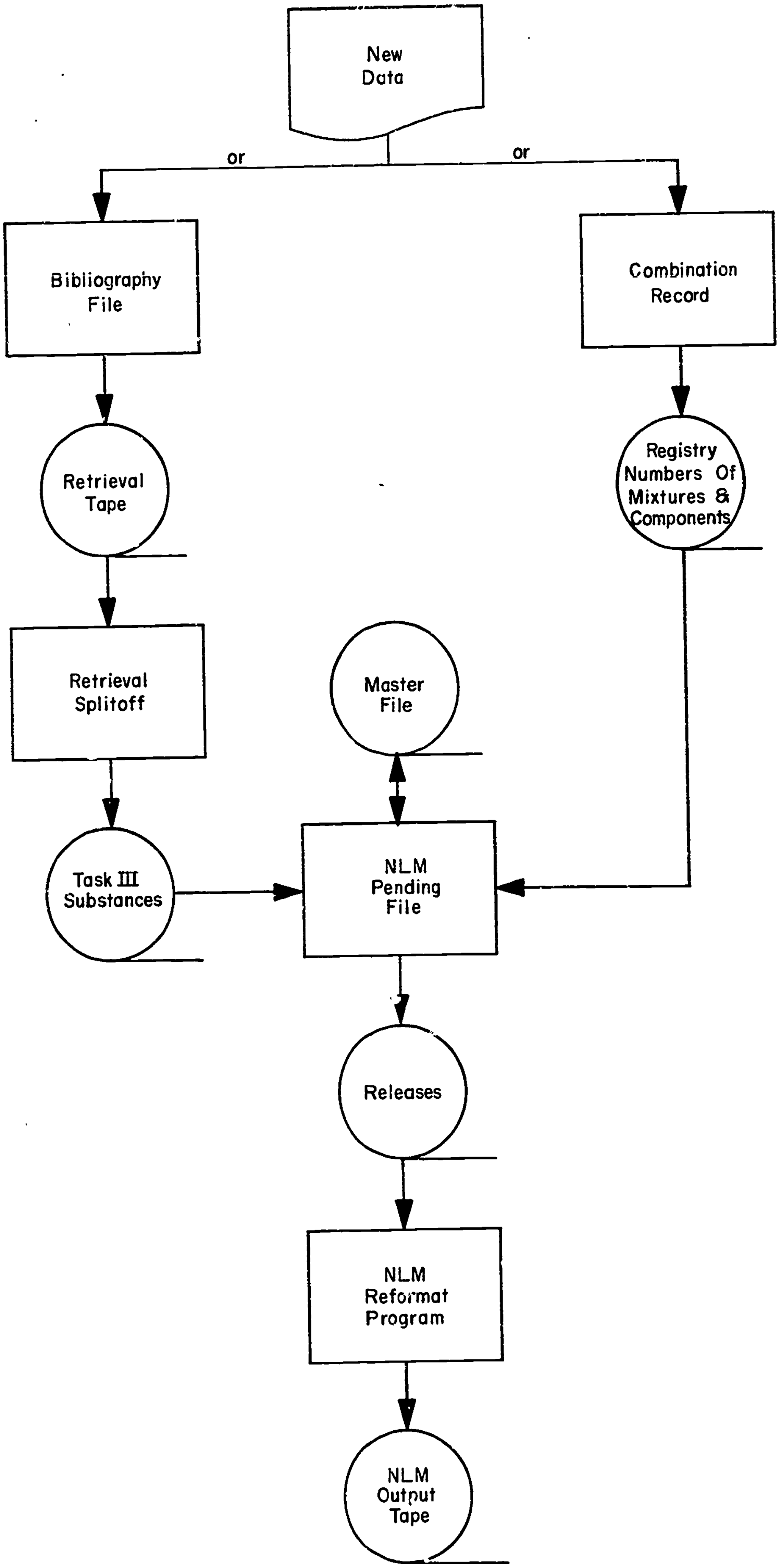


*Figure 2. NLM TAPE OUTPUT SYSTEM*

A mixture is defined as a physical combination of two or more components in which the latter's ratios and identities may or may not be given. Generally, in the case of mixtures, registration is based on the individual compone ts comprising the mixture; mixtures possessing identical components receive the same Registry Number. Each component (except for residues and generically described components such as tars, fluorides, and fatty acids) as well as the mixture itself is assigned a Registry Number. A component may be: 1) a single substance such as a compound, or 2) a mixture which itself is composed of two or more components, or 3) a residue which is either stated or clearly implied. Residues identified only by the term "residues" are not assigned Registry Numbers, nor are such generic terms as "fluorides" or "fatty acids." However, these terms are considered in determining the Registry Number of mixtures. For example, if Mixture A is comprised of Components 1 and 2 and Mixture B of Components 1 and 2 plus a residue, they are assigned different Registry Numbers.

Another illustration of how the above criteria operate can be shown by the registration rationale of mixtures of mixtures. If Mixture 1 is comprised of Compounds A, B, C, and D; and Mixture 2 is comprised of Components 1 and 2; and it is determined that Component 1 is a mixture of Compounds A and B and that Component 2 is a mixture of Compounds C and D, Mixtures 1 and 2 will be assigned different Registry Numbers since Mixture 1 is described as having four components and Mixture 2 is described as having two components. However, the File's cross-referencing feature recognizes that Compounds A, B, C, and D are present in both Mixtures 1 and 2 and in Components 1 and 2 respectively, but it also recognizes that Components 1 and 2 are present in both of the mixtures. (See Figure 3.)

Certain natural origin substances such as oils, concentrates, and juices represent a type of mixture that is not registered on the basis of components.

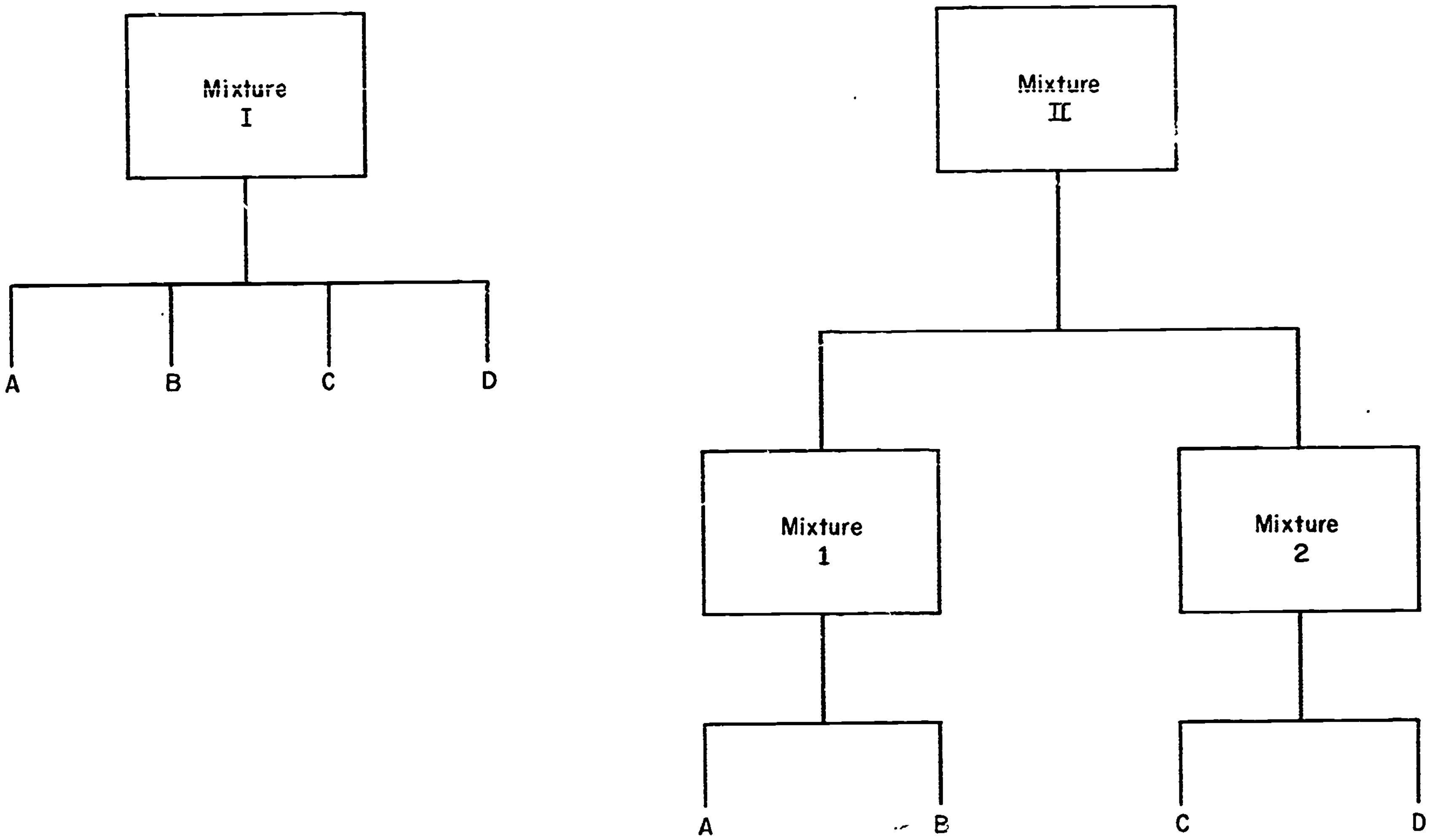


*Figure 3.* EXAMPLE OF MIXTURES THAT RECEIVE DIFFERENT REGISTRY NUMBERS

Because the descriptions of such substances vary substantially from one source to another due to a wide natural variation in composition, the registration of these substances is based on the name of the mixture including botanical or biological identification, if available, rather than the components. If components are known, they are assigned a Registry Number, and as succeeding sources list new components for the mixture, these are added to the record. However, only one Registry Number is assigned to the mixture itself.

The Registry Number assigned to a mixture is comprised of the prefix MX and a seven-digit Registry Number drawn from the eight-million series, for example, MX8000008. The Registry Number for each component of a mixture is cross-referred to the Registry Number of its parent mixture. This feature identifies mixtures containing a given component as well as the components of a given mixture.

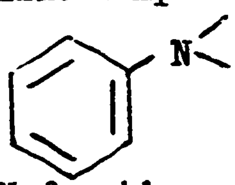
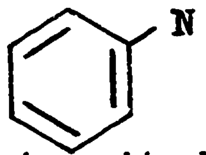
## MeSH CLASS TERM ASSIGNMENT

MeSH Class Terms are generic index headings used in MEDLARS to identify substances containing common structural units. As part of contract NSF-C414, Task III, approximately 500 MeSH Class Terms were defined by NLM and given to CAS. CAS is to assign the appropriate MeSH Class Term(s) to each substance in the Common Data Base for which there is an adequately defined structure on record.

To effect these assignments, the CAS Substructure Search System is used to identify all compounds containing the requisite structural unit while another computer program makes the assignment based upon the substructure search.

The Class Terms originally defined by NLM were, in general, refined by CAS (with NLM approval) to produce search profiles that accurately and precisely defined the structure fragments. In this manner, the total specificity inherent in the Substructure Search System could be used to advantage. However, in the case of some 28 percent of the Class Terms, this was not sufficient. After normal refinement and coding of the Terms, it became evident that a complete redefinition of the Terms was required. This was accomplished by the joint effort of both NLM and CAS staff and unquestionably will result in an enhanced system.

Returning to the general situation; after refinement, profiles were coded for fragment search and, when necessary, for an iterative, atom-by-atom, bond-by-bond search. An example of a term defined by NLM and refined by CAS for searching is as follows:

| NLM | CAS |
|---|---|
| Aniline Compounds | |
|  |  |
| No N-function on N<br>No fused ring<br>No C=O(S) on N | No atom attached to ring or N by a ring bond<br>No N attached to N<br>No =O attached to N<br>No C=O(S) attached to N |

In this example, the principle difference is "No =O attached to N". This was added to preclude nitrobenzene derivatives from being retrieved as aniline compounds.

Although all MeSH Class Terms are generic (by definition), some are broader than others. The various levels of specificity are related through a number of hierarchical series such as the one illustrated below. These series were established principally by NLM, although additional members were added during the refinement of the Class Terms for substructure search.

- Azoles
  - Imidazoles
    - Benzimidazoles
    - Purines
      - Adenines
        - Adenine Nucleotides
      - Guanines
    - Hydantoins
  - Pyrazoles

The program created for Class Term assignment utilizes the hierarchical series in conjunction with substructure search results to assign the appropriate MeSH Class Terms in accordance with the NLM policy of assigning the most specific applicable term to each chemical. For example, the structure for adenosine triphosphate will satisfy the substructure search requests for the following Class Terms:

Azoles
Imidazoles
Purines
Adenines
Adenine Nucleotides
Pyrimidines
Phosphates
Pyrophosphates
Furans
Ethers, Cyclic
Nucleotides
Nucleosides

However, applying the hierarchies after substructure search results in only the Class Terms "Adenine Nucleotides" and "Pyrophosphates" being assigned to the structure for adenosine triphosphate.

For a detailed flow diagram of the complete MeSH Class Term Assignment System, see Figure 4.

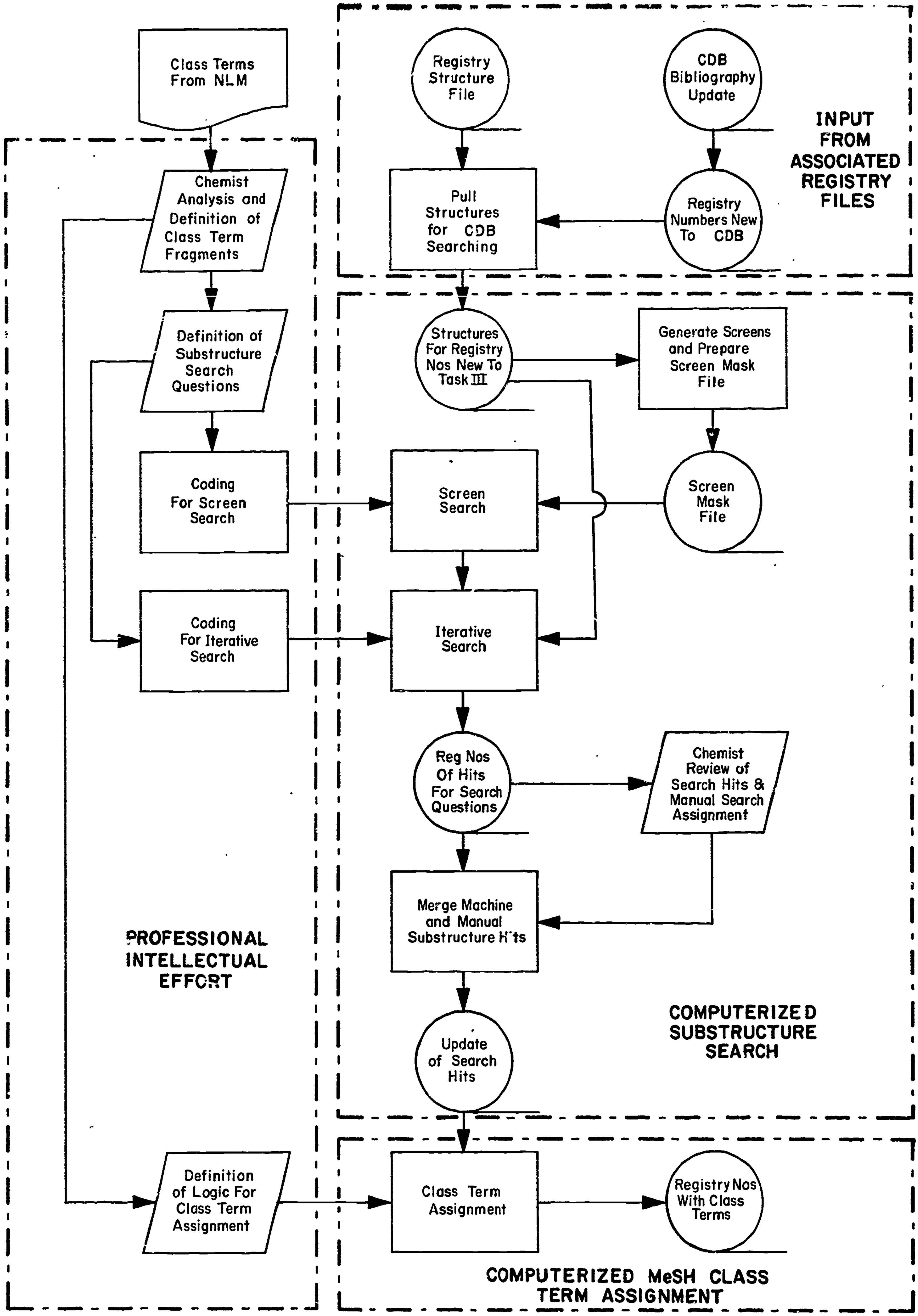

Class Terms From NLM
Chemist Analysis and Definition of Class Term Fragments
Definition of Substructure Search Questions
Coding For Screen Search
Coding For Iterative Search
PROFESSIONAL INTELLECTUAL EFFORT
Definition of Logic For Class Term Assignment
Registry Structure File
CDB Bibliography Update
INPUT FROM ASSOCIATED REGISTRY FILES
Pull Structures for CDB Searching
Registry Numbers New To CDB
Structures For Registry Nos New To Task III
Generate Screens and Prepare Screen Mask File
Screen Search
Screen Mask File
Iterative Search
Reg Nos Of Hits For Search Questions
Chemist Review of Search Hits & Manual Search Assignment
Merge Machine and Manual Substructure Hits
Update of Search Hits
COMPUTERIZED SUBSTRUCTURE SEARCH
Class Term Assignment
Registry Nos With Class Terms
COMPUTERIZED MeSH CLASS TERM ASSIGNMENT

*Figure 4. MeSH CLASS TERM ASSIGNMENT SYSTEM*

# OUTPUT

Three types of output are required by Contract NSF-C414. These are 1) Desktop Analysis Tools (DAT), 2) Manual Structural File, and 3) the Composite File on computer tape. The Composite File for NLM, referred to as the NLM Tape Output System, has already been described in the previous section; details concerning the first two and the Composite File for FDA are presented below.

## DESKTOP ANALYSIS TOOLS

The Task III Desktop Analysis Tool (DAT) is a computer-produced, printed compilation of all compound names contained in the Common Data Base. The DAT is a reference tool that links the Registry Number to the molecular formula and names of the chemical compound or substance. The three specialized DAT's for Common Data Base compounds are: the Name Index in Alphabetical Order, the Name Index in Registry Number Order, and the Molecular Formula Index. The Systematic arrangement of data in each DAT facilitates the handling of chemical information for substances in the Common Data Base.

CAS had previously developed a computerized compound data system under Task I of this contract, which we referred to as a DAT system. However, the requirements for this earlier system were less sophisticated than those of Task III. Therefore, it became necessary for CAS to develop a totally new DAT system to meet the needs of Task III. Two Task III DAT's have been issued to date; the first in February, 1967, the second in May. The February DAT was published principally to establish the future format of each volume and secondarily to test the computer programs required to produce the publication. The May DAT was the first full publication of this type of document for Task III.

Below is a description of each section of the DAT.

## Name Index in Alphabetical Order

This form of the DAT is arranged by the alphabetical order of names (except for laboratory numbers) beginning with the Roman letters constituting the main portion of the name. The following information is provided for each entry:

Registry Number--the unique number given to each substance that identifies that substance throughout all CAS operations.

Item Number--used in conjunction with the Registry Number to provide a unique means of accessing the items of data in the CAS Bibliographic File.

Name Type--identifies the type of name.

Laboratory Numbers--appear in ascending numerical order ahead of the other names which are in alphabetical order.

Molecular Formula--arranged in NOPS sequence.

Source Abbreviation--is printed immediately after the molecular formula.

## Name Index in Registry Number Order

The Name Index in Registry Number Order contains the same information as the Names Indexed in Alphabetical Order. However, as the name implies, this DAT is ordered in ascending Registry Number order. This arrangement permits all names and synonyms associated with a compound or mixture to be grouped together. According to the sorting sequence used, alphabetically prefixed Registry Numbers sort ahead of regular Registry Numbers. Thus, mixtures appear first in this document.

## Molecular Formula Index

This form of DAT is ordered according to the molecular formula arranged in NOPS sequence. The arrangement of this particular volume is influenced by a format convention employed in all the DAT's--the "dot-disconnected" format. Utilized principally to represent the structures and molecular formulas of metal salts of acids, acid salts of bases, quarternary ammonium salts, and addition compounds, this format represents these types of compounds as two or more individual structures separated by a dot. Upon ordering, the format causes the collection of closely related compounds in one place. An example of such a representation is $O_2C_2H_4 \cdot Na$, for sodium acetate. The dot itself appears only on the printed output, not on the machine record, of the molecular formula. In the ordering of the molecular formulas for the DAT's, the portion of the "dot-disconnected" formula that contains the highest carbon count is given first.

Other data given in this DAT include the Registry Number, the Item Number, and the CA index names. Two types of names are listed, the Preferred CA Index Name and the Added CA Index Name. For a detailed flow diagram of the computerized DAT program, see Figure 5. Samples of the results of this program are shown in Figures 6-8.

## MANUAL STRUCTURE FILE

Some 12,000 5 x 8-inch cards comprising the first shipment of the Manual Structure File were delivered in May, 1967. Except for a graphic arts quality, hand-drawn structure, these cards are printed by computer (see Figure 9). In addition to the structure, the file contains the Registry Number, NOPS molecular formula, and CA Preferred Index Name on a large majority of the fully defined structures in the Common Data Base. Two sets of cards were produced, one was ordered by ascending Registry Number and the other by molecular formula. It is anticipated that approximately 8,000 additions will be made to this file during the remainder of the contract.

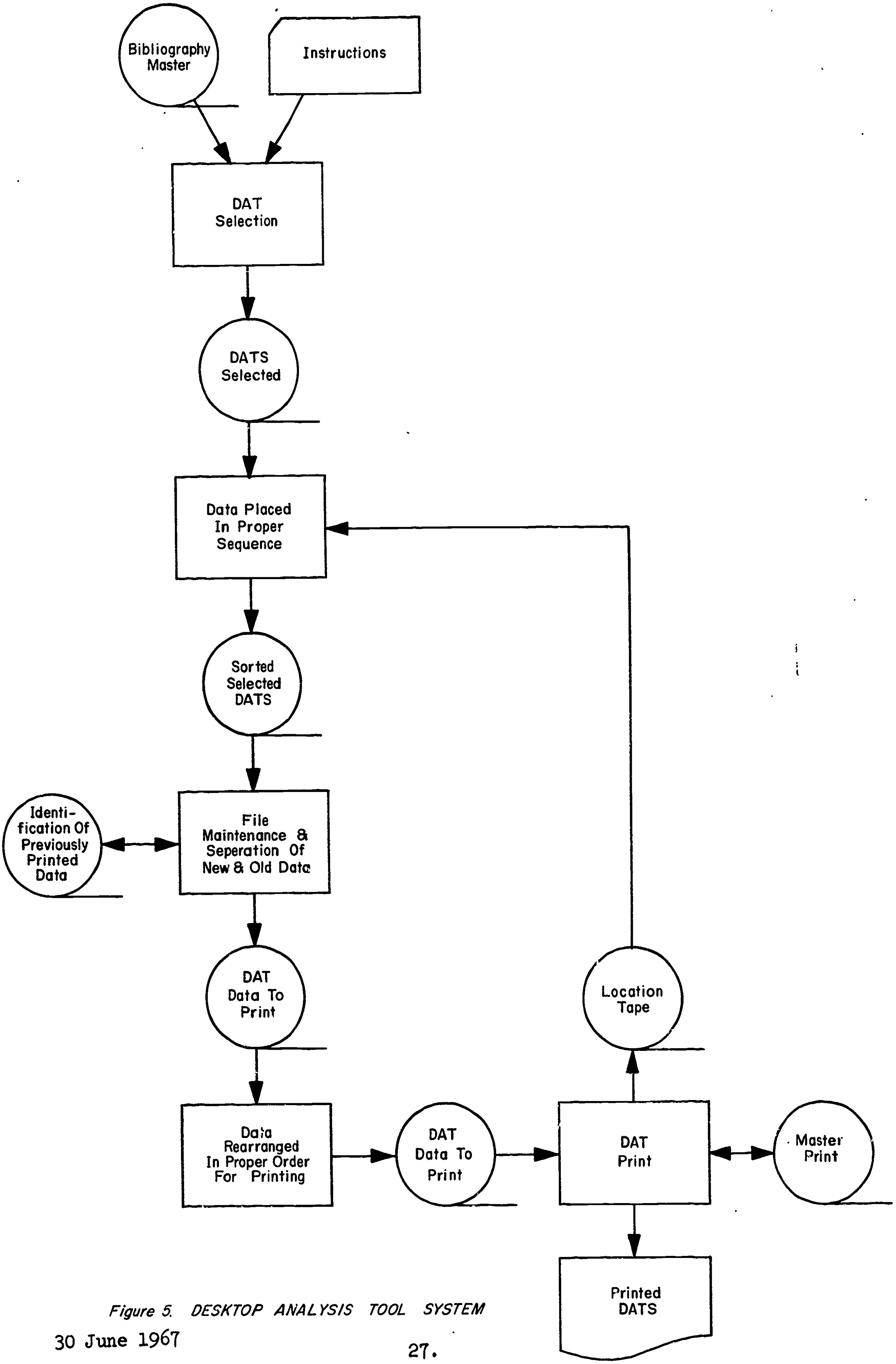
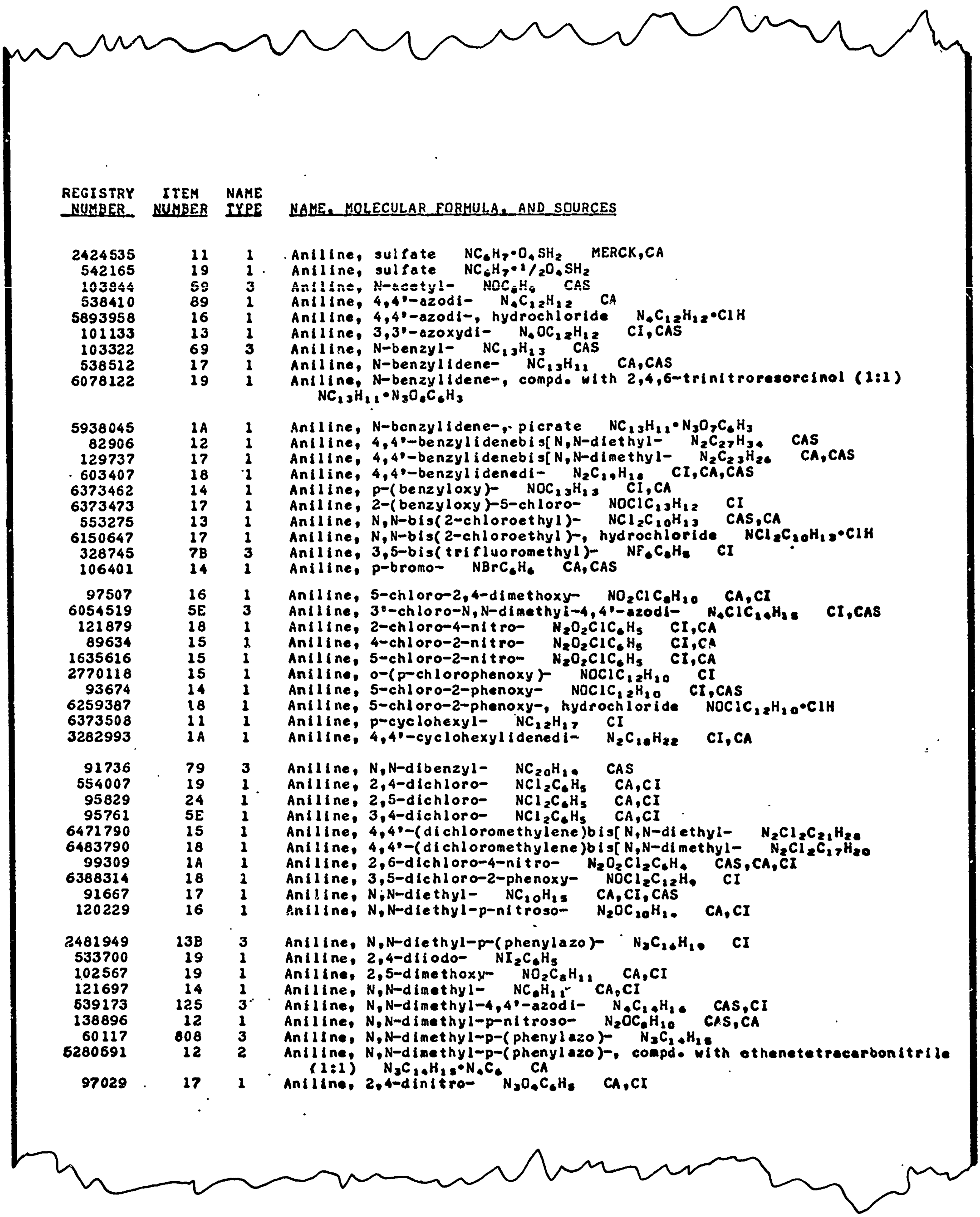
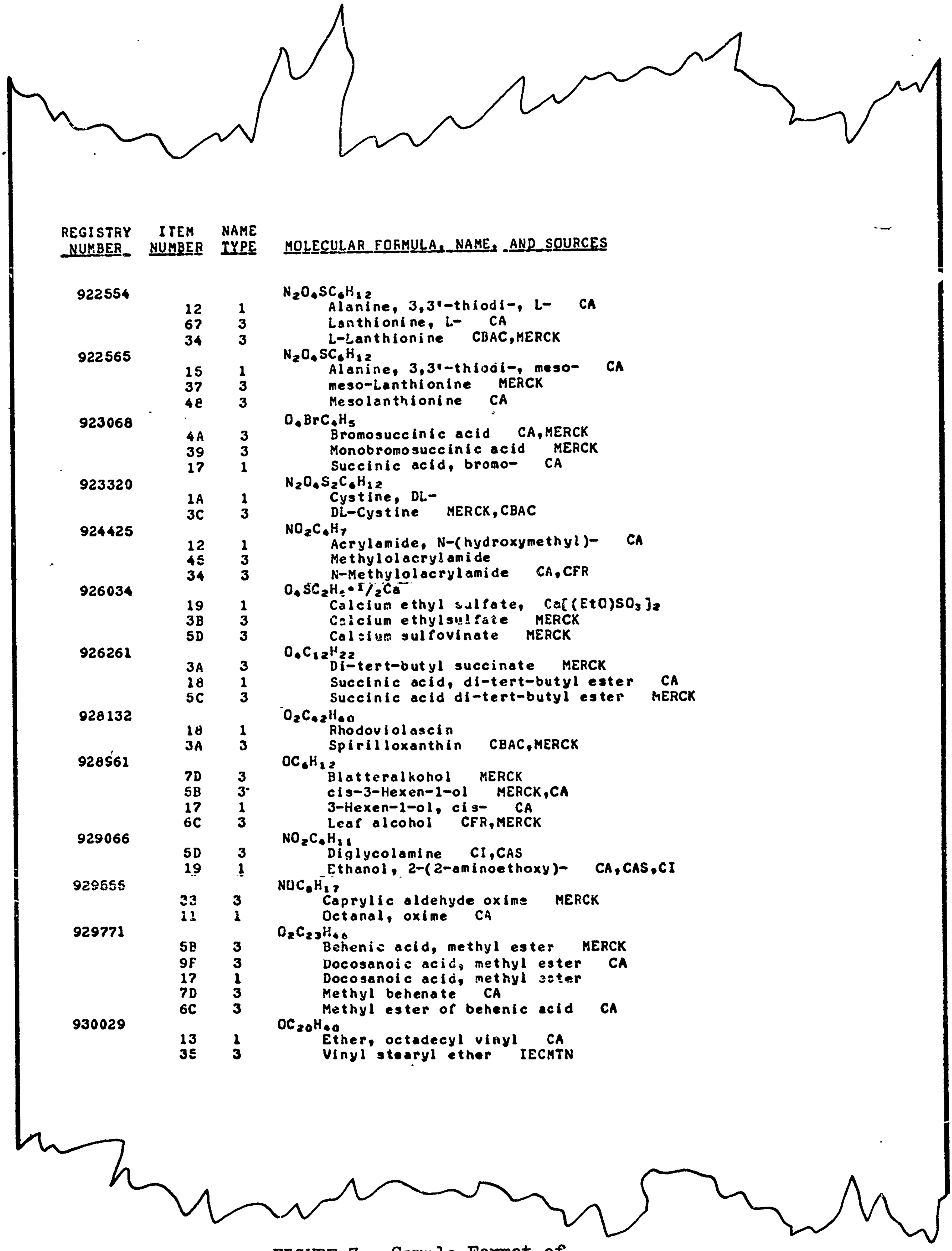

Bibliography Master
Instructions
DAT Selection
DATS Selected
Data Placed In Proper Sequence
Sorted Selected DATS
Identi-fication Of Previously Printed Data
File Maintenance & Seperation Of New & Old Data
DAT Data To Print
Data Rearranged In Proper Order For Printing
DAT Data To Print
DAT Print
Master Print
Location Tape
Printed DATS

*Figure 5. DESKTOP ANALYSIS TOOL SYSTEM*



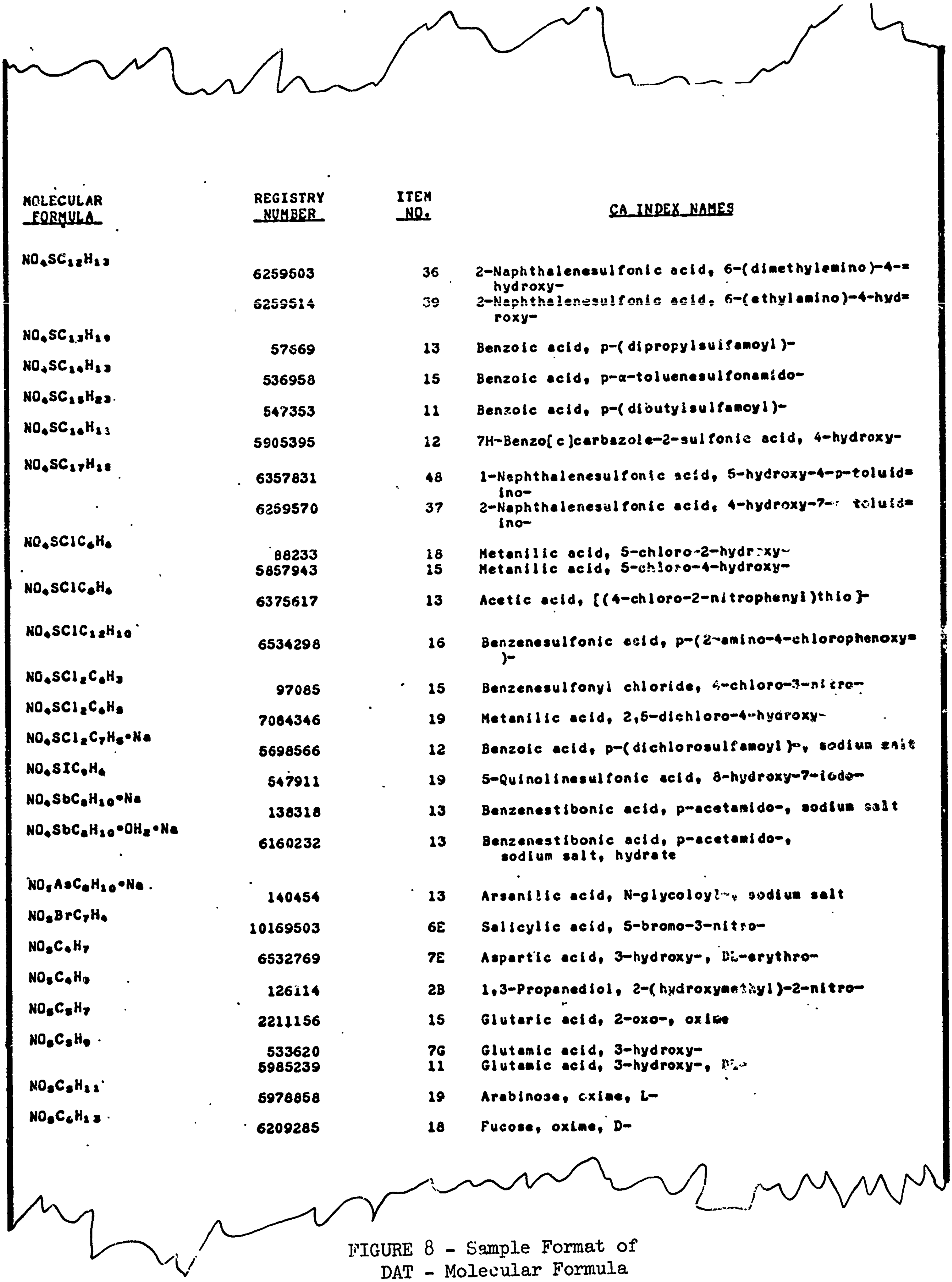
| REGISTRY NUMBER | ITEM NUMBER | NAME TYPE | NAME, MOLECULAR FORMULA, AND SOURCES |
|---|---|---|---|
| 2424535 | 11 | 1 | Aniline, sulfate $NC_6H_7\cdot O_4SH_2$ MERCK,CA |
| 542165 | 19 | 1 | Aniline, sulfate $NC_6H_7\cdot {}^1/_2O_4SH_2$ |
| 103844 | 59 | 3 | Aniline, N-acetyl- $NOC_8H_9$ CAS |
| 538410 | 89 | 1 | Aniline, 4,4'-azodi- $N_4C_{12}H_{12}$ CA |
| 5893958 | 16 | 1 | Aniline, 4,4'-azodi-, hydrochloride $N_4C_{12}H_{12}\cdot ClH$ |
| 101133 | 13 | 1 | Aniline, 3,3'-azoxydi- $N_4OC_{12}H_{12}$ CI,CAS |
| 103322 | 69 | 3 | Aniline, N-benzyl- $NC_{13}H_{13}$ CAS |
| 538512 | 17 | 1 | Aniline, N-benzylidene- $NC_{13}H_{11}$ CA,CAS |
| 6078122 | 19 | 1 | Aniline, N-benzylidene-, compd. with 2,4,6-trinitroresorcinol (1:1) $NC_{13}H_{11}\cdot N_3O_8C_6H_3$ |
| 5938045 | 1A | 1 | Aniline, N-benzylidene-, picrate $NC_{13}H_{11}\cdot N_3O_7C_6H_3$ |
| 82906 | 12 | 1 | Aniline, 4,4'-benzylidenebis[N,N-diethyl- $N_2C_{27}H_{34}$ CAS |
| 129737 | 17 | 1 | Aniline, 4,4'-benzylidenebis[N,N-dimethyl- $N_2C_{23}H_{26}$ CA,CAS |
| 603407 | 18 | 1 | Aniline, 4,4'-benzylidenedi- $N_2C_{19}H_{18}$ CI,CA,CAS |
| 6373462 | 14 | 1 | Aniline, p-(benzyloxy)- $NOC_{13}H_{13}$ CI,CA |
| 6373473 | 17 | 1 | Aniline, 2-(benzyloxy)-5-chloro- $NOClC_{13}H_{12}$ CI |
| 553275 | 13 | 1 | Aniline, N,N-bis(2-chloroethyl)- $NCl_2C_{10}H_{13}$ CAS,CA |
| 6150647 | 17 | 1 | Aniline, N,N-bis(2-chloroethyl)-, hydrochloride $NCl_2C_{10}H_{13}\cdot ClH$ |
| 328745 | 7B | 3 | Aniline, 3,5-bis(trifluoromethyl)- $NF_6C_8H_5$ CI |
| 106401 | 14 | 1 | Aniline, p-bromo- $NBrC_6H_6$ CA,CAS |
| 97507 | 16 | 1 | Aniline, 5-chloro-2,4-dimethoxy- $NO_2ClC_8H_{10}$ CA,CI |
| 6054519 | 5E | 3 | Aniline, 3'-chloro-N,N-dimethyl-4,4'-azodi- $N_4ClC_{14}H_{15}$ CI,CAS |
| 121879 | 18 | 1 | Aniline, 2-chloro-4-nitro- $N_2O_2ClC_6H_5$ CI,CA |
| 89634 | 15 | 1 | Aniline, 4-chloro-2-nitro- $N_2O_2ClC_6H_5$ CI,CA |
| 1635616 | 15 | 1 | Aniline, 5-chloro-2-nitro- $N_2O_2ClC_6H_5$ CI,CA |
| 2770118 | 15 | 1 | Aniline, o-(p-chlorophenoxy)- $NOClC_{12}H_{10}$ CI |
| 93674 | 14 | 1 | Aniline, 5-chloro-2-phenoxy- $NOClC_{12}H_{10}$ CI,CAS |
| 6259387 | 18 | 1 | Aniline, 5-chloro-2-phenoxy-, hydrochloride $NOClC_{12}H_{10}\cdot ClH$ |
| 6373508 | 11 | 1 | Aniline, p-cyclohexyl- $NC_{12}H_{17}$ CI |
| 3282993 | 1A | 1 | Aniline, 4,4'-cyclohexylidenedi- $N_2C_{18}H_{22}$ CI,CA |
| 91736 | 79 | 3 | Aniline, N,N-dibenzyl- $NC_{20}H_{19}$ CAS |
| 554007 | 19 | 1 | Aniline, 2,4-dichloro- $NCl_2C_6H_5$ CA,CI |
| 95829 | 24 | 1 | Aniline, 2,5-dichloro- $NCl_2C_6H_5$ CA,CI |
| 95761 | 5E | 1 | Aniline, 3,4-dichloro- $NCl_2C_6H_5$ CA,CI |
| 6471790 | 15 | 1 | Aniline, 4,4'-(dichloromethylene)bis[N,N-diethyl- $N_2Cl_2C_{21}H_{28}$ |
| 6483790 | 18 | 1 | Aniline, 4,4'-(dichloromethylene)bis[N,N-dimethyl- $N_2Cl_2C_{17}H_{20}$ |
| 99309 | 1A | 1 | Aniline, 2,6-dichloro-4-nitro- $N_2O_2Cl_2C_6H_4$ CAS,CA,CI |
| 6388314 | 18 | 1 | Aniline, 3,5-dichloro-2-phenoxy- $NOCl_2C_{12}H_9$ CI |
| 91667 | 17 | 1 | Aniline, N,N-diethyl- $NC_{10}H_{15}$ CA,CI,CAS |
| 120229 | 16 | 1 | Aniline, N,N-diethyl-p-nitroso- $N_2OC_{10}H_{14}$ CA,CI |
| 2481949 | 13B | 3 | Aniline, N,N-diethyl-p-(phenylazo)- $N_3C_{16}H_{19}$ CI |
| 533700 | 19 | 1 | Aniline, 2,4-diiodo- $NI_2C_6H_5$ |
| 102567 | 19 | 1 | Aniline, 2,5-dimethoxy- $NO_2C_8H_{11}$ CA,CI |
| 121697 | 14 | 1 | Aniline, N,N-dimethyl- $NC_8H_{11}$ CA,CI |
| 539173 | 125 | 3 | Aniline, N,N-dimethyl-4,4'-azodi- $N_4C_{14}H_{16}$ CAS,CI |
| 138896 | 12 | 1 | Aniline, N,N-dimethyl-p-nitroso- $N_2OC_8H_{10}$ CAS,CA |
| 60117 | 808 | 3 | Aniline, N,N-dimethyl-p-(phenylazo)- $N_3C_{14}H_{15}$ |
| 6280691 | 12 | 2 | Aniline, N,N-dimethyl-p-(phenylazo)-, compd. with ethenetetracarbonitrile (1:1) $N_3C_{14}H_{15}\cdot N_4C_6$ CA |
| 97029 | 17 | 1 | Aniline, 2,4-dinitro- $N_3O_4C_6H_5$ CA,CI |

FIGURE 6 - Sample Format of DAT - Name Index in Alphabetical Order



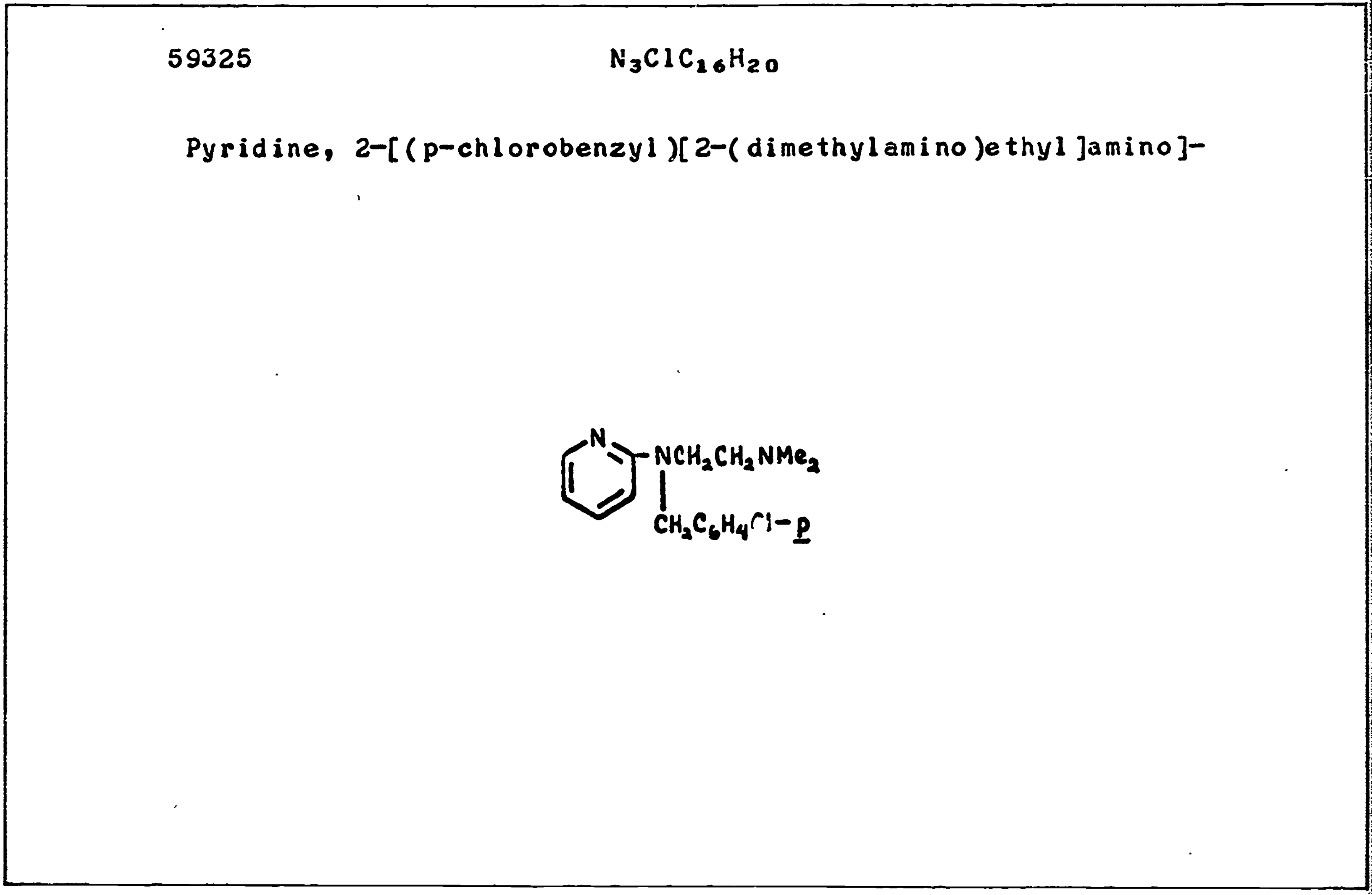
| REGISTRY NUMBER | ITEM NUMBER | NAME TYPE | MOLECULAR FORMULA, NAME, AND SOURCES |
|---|---|---|---|
| 922554 | | | $N_2O_4SC_6H_{12}$ |
| | 12 | 1 | Alanine, 3,3'-thiodi-, L- CA |
| | 67 | 3 | Lanthionine, L- CA |
| | 34 | 3 | L-Lanthionine CBAC,MERCK |
| 922565 | | | $N_2O_4SC_6H_{12}$ |
| | 15 | 1 | Alanine, 3,3'-thiodi-, meso- CA |
| | 37 | 3 | meso-Lanthionine MERCK |
| | 4E | 3 | Mesolanthionine CA |
| 923068 | | | $O_4BrC_4H_5$ |
| | 4A | 3 | Bromosuccinic acid CA,MERCK |
| | 39 | 3 | Monobromosuccinic acid MERCK |
| | 17 | 1 | Succinic acid, bromo- CA |
| 923320 | | | $N_2O_4S_2C_6H_{12}$ |
| | 1A | 1 | Cystine, DL- |
| | 3C | 3 | DL-Cystine MERCK,CBAC |
| 924425 | | | $NO_2C_4H_7$ |
| | 12 | 1 | Acrylamide, N-(hydroxymethyl)- CA |
| | 4E | 3 | Methylolacrylamide |
| | 34 | 3 | N-Methylolacrylamide CA,CFR |
| 926034 | | | $O_4SC_2H_6 \cdot {}^1/_2Ca$ |
| | 19 | 1 | Calcium ethyl sulfate, $Ca[(EtO)SO_3]_2$ |
| | 3B | 3 | Calcium ethylsulfate MERCK |
| | 5D | 3 | Calcium sulfovinate MERCK |
| 926261 | | | $O_4C_{12}H_{22}$ |
| | 3A | 3 | Di-tert-butyl succinate MERCK |
| | 18 | 1 | Succinic acid, di-tert-butyl ester CA |
| | 5C | 3 | Succinic acid di-tert-butyl ester MERCK |
| 928132 | | | $O_2C_{42}H_{60}$ |
| | 18 | 1 | Rhodoviolascin |
| | 3A | 3 | Spirilloxanthin CBAC,MERCK |
| 928561 | | | $OC_6H_{12}$ |
| | 7D | 3 | Blatteralkohol MERCK |
| | 5B | 3 | cis-3-Hexen-1-ol MERCK,CA |
| | 17 | 1 | 3-Hexen-1-ol, cis- CA |
| | 6C | 3 | Leaf alcohol CFR,MERCK |
| 929066 | | | $NO_2C_4H_{11}$ |
| | 5D | 3 | Diglycolamine CI,CAS |
| | 19 | 1 | Ethanol, 2-(2-aminoethoxy)- CA,CAS,CI |
| 929655 | | | $NOC_8H_{17}$ |
| | 33 | 3 | Caprylic aldehyde oxime MERCK |
| | 11 | 1 | Octanal, oxime CA |
| 929771 | | | $O_2C_{23}H_{46}$ |
| | 5B | 3 | Behenic acid, methyl ester MERCK |
| | 9F | 3 | Docosanoic acid, methyl ester CA |
| | 17 | 1 | Docosanoic acid, methyl ester |
| | 7D | 3 | Methyl behenate CA |
| | 6C | 3 | Methyl ester of behenic acid CA |
| 930029 | | | $OC_{20}H_{40}$ |
| | 13 | 1 | Ether, octadecyl vinyl CA |
| | 3E | 3 | Vinyl stearyl ether IECNTN |

FIGURE 7 - Sample Format of DAT - Name Index in Registry Number Order



| MOLECULAR FORMULA | REGISTRY NUMBER | ITEM NO. | CA INDEX NAMES |
|---|---|---|---|
| $NO_4SC_{12}H_{13}$ | 6259503 | 36 | 2-Naphthalenesulfonic acid, 6-(dimethylamino)-4-hydroxy- |
| | 6259514 | 39 | 2-Naphthalenesulfonic acid, 6-(ethylamino)-4-hydroxy- |
| $NO_4SC_{13}H_{19}$ | 57669 | 13 | Benzoic acid, p-(dipropylsulfamoyl)- |
| $NO_4SC_{14}H_{13}$ | 536958 | 15 | Benzoic acid, p-α-toluenesulfonamido- |
| $NO_4SC_{15}H_{23}$ | 547353 | 11 | Benzoic acid, p-(dibutylsulfamoyl)- |
| $NO_4SC_{16}H_{13}$ | 5905395 | 12 | 7H-Benzo[c]carbazole-2-sulfonic acid, 4-hydroxy- |
| $NO_4SC_{17}H_{15}$ | 6357831 | 48 | 1-Naphthalenesulfonic acid, 5-hydroxy-4-p-toluidino- |
| | 6259570 | 37 | 2-Naphthalenesulfonic acid, 4-hydroxy-7-[illegible] toluidino- |
| $NO_4SClC_6H_6$ | 88233 | 18 | Metanilic acid, 5-chloro-2-hydroxy- |
| | 5857943 | 15 | Metanilic acid, 5-chloro-4-hydroxy- |
| $NO_4SClC_8H_6$ | 6375617 | 13 | Acetic acid, [(4-chloro-2-nitrophenyl)thio]- |
| $NO_4SClC_{12}H_{10}$ | 6534298 | 16 | Benzenesulfonic acid, p-(2-amino-4-chlorophenoxy)- |
| $NO_4SCl_2C_6H_3$ | 97085 | 15 | Benzenesulfonyl chloride, 4-chloro-3-nitro- |
| $NO_4SCl_2C_6H_5$ | 7084346 | 19 | Metanilic acid, 2,5-dichloro-4-hydroxy- |
| $NO_4SCl_2C_7H_5 \cdot Na$ | 5698566 | 12 | Benzoic acid, p-(dichlorosulfamoyl)-, sodium salt |
| $NO_4SIC_9H_6$ | 547911 | 19 | 5-Quinolinesulfonic acid, 8-hydroxy-7-iodo- |
| $NO_4SbC_8H_{10} \cdot Na$ | 138318 | 13 | Benzenestibonic acid, p-acetamido-, sodium salt |
| $NO_4SbC_8H_{10} \cdot OH_2 \cdot Na$ | 6160232 | 13 | Benzenestibonic acid, p-acetamido-, sodium salt, hydrate |
| $NO_5AsC_8H_{10} \cdot Na$ | 140454 | 13 | Arsanilic acid, N-glycoloyl-, sodium salt |
| $NO_5BrC_7H_4$ | 10169503 | 6E | Salicylic acid, 5-bromo-3-nitro- |
| $NO_5C_4H_7$ | 6532769 | 7E | Aspartic acid, 3-hydroxy-, DL-erythro- |
| $NO_5C_4H_9$ | 126114 | 2B | 1,3-Propanediol, 2-(hydroxymethyl)-2-nitro- |
| $NO_5C_5H_7$ | 2211156 | 15 | Glutaric acid, 2-oxo-, oxime |
| $NO_5C_5H_9$ | 533620 | 7G | Glutamic acid, 3-hydroxy- |
| | 5985239 | 11 | Glutamic acid, 3-hydroxy-, DL- |
| $NO_5C_5H_{11}$ | 5978858 | 19 | Arabinose, oxime, L- |
| $NO_5C_6H_{13}$ | 6209285 | 18 | Fucose, oxime, D- |

FIGURE 8 - Sample Format of DAT - Molecular Formula Index



59325 $N_3ClC_{16}H_{20}$

Pyridine, 2-[(p-chlorobenzyl)[2-(dimethylamino)ethyl]amino]-

N
$NCH_2CH_2NMe_2$
$CH_2C_6H_4Cl$-p

FIGURE 9. Example of Manual Structure File Card

## TAPE OUTPUT

Contract C414 requires CAS to develop the computer routines needed to convert the output of its own computer, an IBM 360, to a form compatible with the NLM computer, a Honeywell 200. The method being developed by CAS to meet this requirement was discussed previously. To acquire this capability has necessitated a very close liaison between personnel of both organizations. Seven test tapes have been generated by CAS and tested by NLM in an effort to establish the format and language required by the Honeywell computer. The seventh tape which was sent to NLM in June of this year has shown that the development of this format is near success. By making modifications to tape No. 7, it is expected that the next test tape will successfully perform its function. After these difficulties have been overcome, the Composite File will be put on computer tape and sent to NLM. See Appendix D for a brief description of the contents of each test tape.

## FDA COMPOSITE FILE

The FDA Composite File is a compilation of the data contained in the Common Data Base. In this respect, it is identical to the NLM Tape Output previously described. However, a major difference exists between the two computer files. While the NLM file is to be in a format readable by a Honeywell 200 computer, the FDA file will be readable by an IBM 360 computer, the computer for which CAS programs are being reprogrammed.

At FDA's request, the delivery of the tapes, documentation and programs needed to process the Composite File has been delayed pending the installation of FDA's computer and the complete reprogramming of CAS' routines from the IBM 7010 computer to the IBM 360 computer.

# FDA REGISTRY WORKSHOP

FDA's evolving computerized information system will interface directly with the CAS Chemical Compound Registry System. For this reason, the FDA requested and CAS conducted a training session for a group of FDA personnel to familiarize them with the Registry System.

The symposium was held at FDA training headquarters in Arlington, Virginia and was attended by 14 FDA chemists drawn from several of that organization's divisions and bureaus. The primary interests of these professionals lay in the following fields:

- Drug analysis
- Analytical chemistry
- Food chemistry
- Microbiology of antibiotics
- Pharmaceutical chemistry
- Biochemistry
- Toxicology
- Science information retrieval

The CAS personnel who acted as instructors were the following:

W. C. Davenport--Senior Staff Advisor

M. K. Park--Head, Formula Indexing Department

R. E. Stobaugh--Technical Advisor, Chemical Compound Registry Division

R. W. White--Assistant to the Director, Research and Development Division

Although the original purpose of the workshop was to present an overview of the Chemical Compound Registry System, the actual workshop resulted in a fairly comprehensive study of the procedures and conventions used in the Registry System, primarily due to the study's reception by FDA personnel and the quality of the instruction.

During the three-day symposium, many specific areas of the Chemical Compound Registry System were discussed, including the following:

1. Overview of CAS computer-base information system
2. Confidentiality of data
3. Overview of the Registry System
4. General structure conventions
5. Special structure conventions with particular emphasis upon stereochemistry
6. Handling of nomenclature (names, DAT's, name-match, cross-references)
7. Methods of manual registration
8. Substructure search methods

During these discussions, the instructional materials outlined in Table II were used to aid in the recipient's understanding of the subject matter and to act as reference material when the symposium ended. In addition to the presentations given by the CAS personnel, work sessions were held to give the FDA chemists first-hand experience in some of the techniques used at CAS. Special emphasis was placed upon the use of DAT's.

Communications received from FDA after the session indicated that the workshop was very helpful to the FDA chemists and that their knowledge of the Registry System had been greatly enhanced by the three-day symposium. Interest was such that many of the attendants requested that additional training sessions be held.

Many other meetings and conferences were held between personnel of the various organizations involved in Contract C414. A list of these, including a brief description of the topics discussed, is given in Table III.

TABLE II

## INSTRUCTIONAL MATERIAL FOR FDA WORKSHOP

1. Registry System Description

   A very general description of the system and the fundamentals, including the work flow and description of the Registry form.

2. Registry System Stereochemistry

   Treatment of general and special text descriptors, including the list of permitted descriptors.

3. Registry System Structure Conventions

   Conventions in general and for metal salts, addition compounds, salts of bases, quaternary compounds, carbohydrates, incompletely-described structures, metal coordination compounds, and inorganic compounds.

4. Registry System Manual Registration

   General description, descriptions of mixture and proposed structure handling.

5. The Naming and Indexing of Chemical Compounds from Chemical Abstracts.

6. Registry Sheet forms.

7. Reprints of papers:

   a. F. Tate, Progress Toward A Computer-Based Chemical Information System, C & EN, January 23, 1967.

   b. D. Leiter, H. Morgan, and R. Stobaugh, Installation and Operation of a Registry For Chemical Compounds, J. Chem. Doc. Volume 5 No. 238 (1965)

   c. H. Morgan, Generation of Unique Machine Description For Chemical Structures, J. Chem. Doc. Volume 5 No. 107 (1965).

   d. D. Whittingham, F. Wetsel, and H. Morgan, Computer-Based Subject Index Support System, J. Chem. Doc. Volume 6 No. 4 (1966).

TABLE III

SIGNIFICANT CONFERENCES HELD BETWEEN PARTIES
OF CONTRACT NSF-C414, TASK III, FROM 1 JULY 1966 THROUGH 30 JUNE 1967

| Date | Organizations Represented | Held At | Purpose |
|---|---|---|---|
| 20 July 1966 | NLM, CAS | NLM | To begin to establish systems specification and basic record formats. |
| 16 Aug. 1966 | NLM, FDA, NSF, CAS | CAS | General discussions relating to technical and administrative requirements of contract. |
| 7 Sept. 1966 | NLM, CAS | NLM | Technical discussions relating to 1) definition of MeSH Class Terms used by NLM, and 2) problems arising from samples of MeSH Class Terms sent to CAS by NLM. |
| 21 Oct. 1966 | NLM, CAS | CAS | Technical discussions relating to 1) Composite File format, 2) CAS Sort Key, and 3) Combination File |
| 17 Jan. 1967 | NLM, CAS | NLM | To discuss programming problems related to the Composite File and to acquaint CAS personnel with NLM and the uses to which Task III data will be put. |
| 23 Feb. 1967 | NLM, CAS | NLM | Resolution of technical problems related to MeSH Class Term assignment and substructure search and to review philosophy of NLM indexing system used in MEDLARS and Index Medicus. |
| 24 Feb. 1967 | FDA, CAS | CAS | To discuss 1) CAS Substructure Search System, 2) training of FDA personnel in the Registry System, 3) status report of 10 Feb. 1967, 4) confidentiality of FDA data. |
| 20-22 March 1967 | FDA, CAS | FDA | In depth workshop on Registry System including 1) use of DAT's, 2) substructure search, 3) manual registration, 4) structure conventions, and 5) nomenclature. |
| 24 April 1967 | FDA, CAS | FDA | General technical considerations relating to Task III extension. |
| 4 May 1967 | NLM, CAS | NLM | Clarification of NLM needs relating to Task III extension. |
| 10 May 1967 | FDA, CAS | FDA | Budget considerations relating to Task III extension. |
| 23 May 1967 | FDA, CAS | CAS | Review of CAS Registry System. |

# APPENDIXES

30 June 1967

# APPENDIX A

# GLOSSARY

30 June 1967

# GLOSSARY

CA - *Chemical Abstracts*

CAS - Chemical Abstracts Service

Combination - This is a mixture in which all the components are identified and in which the ratio of components may or may not be specified. A Combination may also be defined only in terms of the "so-called" active components, and the ratio of the active components may or may not be given.

Common Data Base - This is made up of the published data associated with the substances specified in Task III for processing under this contract. The data elements in the Common Data Base are: (1) whatever structural description of the substance is provided; (2) Inverted Molecular Formula, when available; (3) Nomenclature; (4) Source Codes; (5) *CA* references for those synonyms taken only from the Registry files.

Composite File - A magnetic tape file which includes the data associated with the substances in the Common Data Base and associated MeSH Class Terms. The File will be arranged in name order and will include for each substance the following data as they are available:

1. Registry Number
2. Nomenclature
3. Name-Type Code
4. CAS Item Number
5. MeSH Terms
6. Source Codes
7. Inverted Molecular Formula

Compound - A single substance made up of identical molecular species.

Connection Table - Atom-by-atom inventory of the molecule which shows each atom, the atoms connected directly to it, and types of linking bonds. Mass number, coordination number valence, and charges are shown whenever they are required for exact identification. Stereochemical data are included.

Desktop Analysis Tools - Printed lists of specially selected substances for use in the analysis of input information to help identify already registered substances. These tools will include listings of names, laboratory numbers and acronyms with the associated Registry Numbers. Molecular formula indexes are also included.

FDA - Food and Drug Administration

Index Name - A name of a substance as used in the current Subject and Formula Indexes to *Chemical Abstracts*. The Preferred Index Name is used as the main heading in the Subject Index for index entries associated with the substance; this name is also used as the Formula Index entry for a compound. Added Index Names are special-purpose entries in the Subject Index which emphasize special structural features or identify useful author-derived nomenclature.

Inverted Molecular Formula - In this form of molecular formula, nitrogen, oxygen, phosphorus and sulfur are listed first in that order; these are followed alphabetically by symbols for elements other than carbon and hydrogen which come last.

Item Number - A five-digit number that, used in conjunction with a Registry Number, identifies each item of data (e.g., each name, structure, molecular formula, etc.) associated with that Registry Number.

MEDLARS - Medical Literature Analysis and Retrieval System

M^SH Term - This is the Medical Subject Heading. In the context of this contract a MeSH Term is either an assigned specific name (Unique MeSH Term) for a substance or a generic characteristic (MeSH Class Heading) related to the substance to which the Class Term is assigned. A given substance may have both a Unique and a Class MeSH Term.

Unique MeSH Term - Not all substances are represented by a Unique MeSH Term. The initial assignment of a Unique MeSH Term is always carried out by the NLM staff. Once a substance has been assigned a Unique MeSH Term, this Term is always used within the MEDLARS System as an index entry for the substance. Presently NLM has between 1,000 and 1,200 Unique MeSH Terms corresponding to substances. These Terms are a form of nomencla ure in CAS Registry files.

MeSH Class Term - This is a generic heading representing a family of substances in MEDLARS. Certain Class Terms are based on substructural units, for example, phenothiazines, while others are based on biological activity, etc., for example, antibiotics.

Molecular Formula - A listing of the type and total number of each atom present in a molecule.

Name-Type Code - This is a numerical code attached to each entry in the Registry Nomenclature File. The codes identify:

1. Preferred CA Index Name (Inverted Form)
2. Added CA Index Name
3. Author or Trivial Name
4. Preferred CA Index Name (Uninverted Form)

NLM - National Library of Medicine

Nomenclature - All names for substances including acronyms and laboratory numbers.

NSF - National Science Foundation

Organic Compounds - For the purposes of this contract these are carbon-containing compounds.

Registration - The process of determining the existence or nonexistence of a substance in the Registry Files. The process includes the assignment of a unique number (Registry Number) to each substance that is new to the files; this number is to be used in a large, multifaceted system to associate data related to that substance.

Registry Number - The unique number which is assigned to each substance when it first enters the Registry and which is recalled each time that substance is checked against the file. The Registry Number may be used to identify fully the substance, and in the future it can be used as the address in specialized subject files to identify data associated with the substance. A Registry Number may include alphabetic characters, and will include a computed check digit.

Registry System - The interrelated set of files directly associated with registration and the processes (including manual and computer-based facets) for accomplishing registration. These computer files include any available structural record, the molecular form (when available), nomenclature, and bibliographic data.

Source Code - A set of codes attached to each name in the Registry Nomenclature File that identifies the source(s) in which the name is used, for example, Journal of Biological Chemistry, Chemical Abstracts, and Merck Index, or private sources.

Structural Formula - A projected two-dimensional graphic representation of the atoms and bonds of a molecule.

Substructure - A specified set of atoms interconnected in a specified way; this constellation normally represents less than a complete molecule.

## APPENDIX B

## SOURCES OF THE COMMON DATA BASE

30 June 1967

# SOURCES OF THE COMMON DATA BASE

This Appendix is an alphabetical listing of the books and other references from which data on Common Data Base substances were taken.

*Wilson, C. O., Jones, T. E.: American Drug Index. Philadelphia: J. B. Lippincott Co., 1966.

Council of the Pharmaceutical Society of Great Britain: British Veterinary Codex. 1966 Edition. London: The Pharmaceutical Press. 1966.

Gleason, M. N., Gosselin, R. E., Hodge, H. C.: Clinical Toxicology of Commerical Products. 2nd Edition. Baltimore: Williams and Wilkins Co., 1963. Ingredients Index, pp. 1-126.

*Code of Federal Regulations. Title 21, Chapter 1, (Parts 1-129), Parts 8-9.440, 120, 121.200-121.265. Subject Index to Part 121 and selected portions relating to Part 120 issued since 1 January 1965. Washington, D. C.: U. S. Government Printing Office. 1966. pp. 75-126, 327-367, 400-469. Only substances for which names appear as boldface entries preceded by paragraph numbers were input.

*Code of Federal Regulations. Title 21, Chapter 1, (Parts 130-end), Parts 130-133, 141a-e, 146a-e, 148a-x. Washington, D. C.: U. S. Government Printing Office. 1966. pp. 5-60, 72-223, 261-566, 577-728. Only substances for which names appear as boldface entries preceded by paragraph numbers were input.

The Society of Dyers and Colourists: Colour Index. 2nd Edition. London: Percy Lund, Humphries and Co., Ltd. 1956. Vol. 3; 1963 Suppl., pp. S621-S1124.

International Organization for Standardization: ISO Recommendations, R116, R219, R258, R290, "Common Names for Pesticides." Geneva: ISO. 1959, 1961, 1962, 1963.

Sax, N. I.: Dangerous Properties of Industrial Materials. New York: Reinhold Publishing Corp. 1961.

*McCutcheon, J. W.: Detergents and Emulsifiers. 1966 Edition. Morristown, N. J.: John W. McCutcheon, Inc. 1966.

*Drug and Cosmetic Industry: Drug and Cosmetic Catalog. 17th Edition. New York: Drug and Cosmetic Industry, 1966-67.

Chemical Abstracts Service: Drug File (internal)

*Source is an approved substitution of a newer edition for the one called for in the original contract.

*Farm Chemicals: Farm Chemicals Handbook. 52nd Edition. Willoughby, Ohio: Meister Publishing Co. 1966. pp. D247-99.

Animal Health Institute: Feed Additive Compendium. Minneapolis: Miller Publishing Co. 1966. Section 4, pp. 99-371. (Only substances for which names appear as boldface entries were input.)

National Academy of Sciences-National Research Council: Food Chemicals Codex. (NAS-NRC Publication No. 1143). Washington, D. C.: U. S. Govt. Printing Office. 1963. Parts I-X, pp. 1-722.

Martin, H.: Guide to Chemicals Used in Crop Protection. 4th Edition. London, Ontario: Research Branch, Canada Dept. of Agriculture. 1961.

Spencer, E. Y.: Guide to Chemicals Used in Crop Protection. Supplement to 4th Edition. London, Ontario: Research Branch, Canada Dept. of Agriculture. 1964.

Zimmerman, O. T., Lavine, I.: Handbook of Material Trade Names. Dover, N. H.: Industrial Research Service, Inc. 1953: Supplement I. 1956; Supplement II. 1957; Supplement III. 1960.

Spector, W. C. (Editor): Handbook of Toxicology. Vol. 2, Antibiotics. Philadelphia: W. B. Saund.rs Co. 1957.

Wessel, C. J., Bejuki, W. M.: "Industrial Fungicides." Ind. Eng. Chem. Vol. 51(4), 52A-63A (April, 1959).

de Navarre, Maison G.: International Encyclopedia of Cosmetic Material Trade Names. New York: Moore Publishing Co., Inc. 1957. pp. 3-290.

World Health Organization: International Non-proprietary Names. Cumulative List, 1962. Geneva: WHO. 1962. pp. 7-49.

World Health Organization: International Pharmacopeia. 1st Edition. Geneva: WHO. 1951. Vol. 1. pp. 9-258. Vol. 2. 1955. pp. 3-217. Supplement 1959. pp. 3-106. (Only substances for which names appear in the large, boldface headings were input. Entries differing only in physical state were not separately identified, for example, Sodium Phenobarbital and Sodium Phenobarbital Injection.)

List of Colors Appendix pp. 44-6.

Stecher, P. G. (Editor): The Merck Index of Chemicals and Drugs. 7th Edition. Rahway, N. J.: Merck and Co., Inc. 1960. pp. 1-1121.

The Merck Veterinary Manual: Rahway, N. J.: Merck and Co., Inc. 1955.

*Source is an approved substitution of a newer edition for the one called for in the original contract.

Goodhart, R. S. (Editor): Modern Drug Encyclopedia and Therapeutic Index. 10th Edition. New York: Reuben H. Donnelley Corp. 1965.

Modern Veterinary Practice. Red Book Edition. Vol. 47 (5). (April 15, 1966). pp. 195-263.

Woggan, G. N. (Editor): Mycotoxins in Foodstuffs. Cambridge, Mass.: MIT Press. 1965. pp. 29, 34, 40, 41, 59, 64, 83, 84, 118-121, 127, 140, 177, 266-272.

Committee on National Formulary: The National Formulary. 12th Edition. Washington, D. C.: American Pharmaceutical Association. 1965. pp. 10-428. (Only substances for which names appear in the large, boldface headings were input. Entries differing only in physical state were not separately identified, for example, Sodium Phenobarbital and Sodium Phenobarbital Injection.)

*American Medical Association. New Drugs: 1966 Edition. Chicago: AMA. 1966. pp. 1-543. (Only substances for which names appear in the large, boldface headings were input.)

Poucher, W. A. Perfumes, Cosmetics and Soaps, Vol. 1. 6th Edition. London: Chapman and Hall Ltd. 1959. pp. 3-434. (Entries relating to species of the plant and animal kingdom were not input, however, extracts derived from plants and animals were input.)

Frear, D. E. H. (Editor): Pesticide Index. 3rd Edition. State College, Penna.: College Science Publishers. 1965.

U. S. Pharmacopeial Convention, Inc.: The Pharmacopeia of the United States of America. 17th Revision. New York: U. S. Pharmacopeial Convention, Inc. 1965. pp. 13-766. (Only substances for which names appear in the large, boldface headings were input. Entries differing only in physical state were not separately identified, for example, Sodium Phenobarbital and Sodium Phenobarbital Injection.)

*Folsom, J. Paul (Editor) Physicians' Desk Reference 20th Edition. Oradell, N. J.: Medical Economics, Inc. 1966. pp. 189-273, 502-1092.

Johnson, O. H., Krog, N. E., Poland, J. L.: "Pesticides, Part 1." Chem. Week, Vol. 92, pp. 128-48 (May 25, 1963).

Johnson, O. H., Krog, N. E., Poland, J. L.: "Pesticides, Part 2." Chem. Week, Vol. 92, pp. 63-90 (June 1, 1963).

South African Med. J. Vol. 39, pp. 762-4.

Lehman: Summaries of Pesticide Toxicity. 1965.

*Source is an approved substitution of a newer edition than the one called for in the original contract.

*USAN Council: United States Adopted Names. 4th Edition. New York: U.S. Pharmacopeial Convention, Inc. January, 1966. pp. 9-78.

Unlisted Drugs. Vol. 1 (1948) - Vol. 18 (1) (1966). New York: Special Libraries Association.

U. S. Dept. of Agriculture, Pesticide Regulation Division: USDA Summary of Registered Agricultural Pesticide Chemical Uses. 2nd Edition, Supplement I. Washington, D. C. USDA, Agricultural Research Service. 1965.

*Stephenson, H. C. (Editor): Veterinarians' Blue Book. 14th Edition. New York: Reuben H. Donnelley Corp. 1966. pp. 1-109.

Jones, L. M.: Veterinary Pharmacology and Therapeutics. 3rd Edition: Ames, Iowa; Iowa State Univ. Press. 1965.

*Source is an approved substitution of a newer edition than the one called for in the original contract.

# APPENDIX C

# DELIVERIES

30 June 1967

DELIVERIES

| Item | Date | Recipient | Remarks |
|---|---|---|---|
| Test Tape No. 1 | 23 Sept. 1966 | NLM | 1. Contained 15 artificial examples of Composite File entries.<br>2. Documentation accompanied tape. |
| Test Tape No. 2 | 11 Oct. 1966 | NLM | 1. Contained same examples as No. 1.<br>2. MeSH Term records modified to carry CAS Sort Key.<br>3. Corrections of Tape No. 1 problems made. |
| Test Tape No. 3 | 19 Dec. 1966 | NLM | 1. Contained approximately 200/live entries of the Composite File. |
| Test Tape No. 4 | 5 Jan. 1967 | NLM | 1. Contained change in molecular formula format.<br>2. Correction of Sort Key problem made.<br>3. Description of new molecular formula format and a hexidecimal dump of the test tape provided. |
| Test Tape No. 5 | 11 Jan. 1967 | NLM | 1. Contained modified header labels and MeSH Term records.<br>2. Hexidecimal dump of test tape provided. |
| Test Tape No. 6 | 15 Feb. 1967 | NLM | 1. Contained MeSH Class Terms.<br>2. Hexidecimal dump of tape provided. |
| Test Tape No. 7 | 26 June 1967 | NLM | 1. Minor problems of Test Tape No. 6 resolved.<br>2. Larger number of entries provided. |
| Desktop Analysis Tool, Vol. 1, Section 1 | 28 Jan. 1967 | NLM, FDA, NSF | 1. Published to establish format and test computer routines. |
| Desktop Analysis Tool, Vol. 1, Sections 2 and 3 | 1 Feb. 1967 | NLM, FDA, NSF | |

DELIVERIES (Continued)

| Item | Date | Recipient | Remarks |
|---|---|---|---|
| Desktop Analysis Tool, Edition 2, Sections 1 and 3 | 18 May 1967 | NLM, FDA, NSF | 1. First DAT to contain all available data. |
| Desktop Analysis Tool, Edition 2, Section 2 | 1 June 1967 | NLM, FDA, NSF | |
| Manual Structure File | 19 June 1967 | NLM, FDA, NSF | 1. Initial delivery. |