(1) DN

AD

A CHEMICAL INFORMATION AND DATA SYSTEM

Semi-annual Report

by

Clarence T. Van Meter

Ruth V. Powers

Morris Plotkin

Helen N. Hill

28 February 1969

DEPARTMENT OF THE ARMY
EDGEWOOD ARSENAL
Technical Support Directorate
Technical Data & Value Engineering Management Office
Edgewood Arsenal, Maryland 21010

Contract DAAA15-69-C-0140

UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA PENNSYLVANIA 19104

20

A CHEMICAL INFORMATION AND DATA SYSTEM

Semi-annual Report

by

Clarence T. Van Meter

Ruth V. Powers

Morris Plotkin

Helen N. Hill

28 February 1969

DEPARTMENT OF THE ARMY
EDGEWOOD ARSENAL

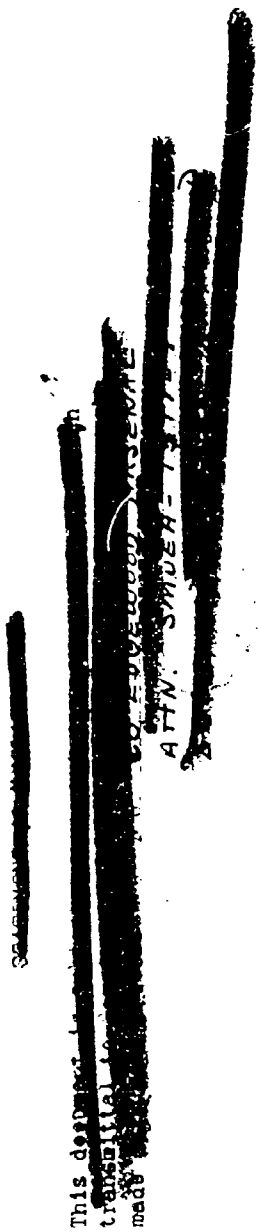
Technical Support Directorate
Technical Data & Value Engineering Management Office
Edgewood Arsenal, Maryland 21010

Contract DAAA15-69-C-0140

Task 2P062101A72702

UNIVERSITY OF PENNSYLVANIA

Philadelphia, Pennsylvania 19104



FOREWORD

The work described in this report was authorized under Task 2PO62101A72702, Army Chemical Information and Data Systems (U). The work was started in July 1964 and is continuing. The information contained in this report represents work accomplished during the period 1 October 1968 - 28 February 1969.

The information in this document has not been cleared for release to the general public.

Acknowledgment

The project is pleased to acknowledge the generous cooperation of members of the staff of the Technical Data Coordination Office and the EDP Systems personnel, Edgewood Arsenal in the conduct of various phases of the work.

Reproduction

DIGEST

This document describes the research and development activities conducted on Project CIDS of the University of Pennsylvania during the period 1 October 1968 - 28 February 1969. All of these activities are pertinent to the creation of a model operational system scheduled for demonstration purposes during the summer and fall of 1969. The content of chemical search screens has been revised in accord with the results of large scale exercising of the earlier experimental system, and one technique for incorporating nonstructural information and data has been devised and is being explored.

Various other improvements in search techniques have been effected, most particularly, file compaction through use of the mechanical chemical code as a representation of the node-connector table and development of a greatly improved atom-by-atom search program. The multi-terminal real-time system has been in operation for a few months and provides the capability of dialing in queries concurrently from four terminals. Documentation with respect to its design, implementation, and use has been prepared. Experimental work on the cathode ray tube as a CIDS input-output device is impressive and an interim report describing the results to date is available.

File construction for the model operational system continues. The total file is expected to approximate 40,000 compounds all of which will be amenable to structural search and a few thousand of which will be searchable separately in terms of nonstructural descriptors provided by Edgewood Arsenal.

The formal CIDS No. 6 Report, which constitutes the Final Report for Contract DA18-035-AMC-288(A) and has just been distributed, documents all chemical search components currently admitted to the revised system. It is a desk-top tool for use in the intellectual assignment of chemical search screens to queries. The next report in the series will describe the retrieval language, the details of search strategy, and the formulation of fully encoded queries. It is currently in an early stage of draft.

TABLE OF CONTENTS

LIST OF TABLES

1. Introduction

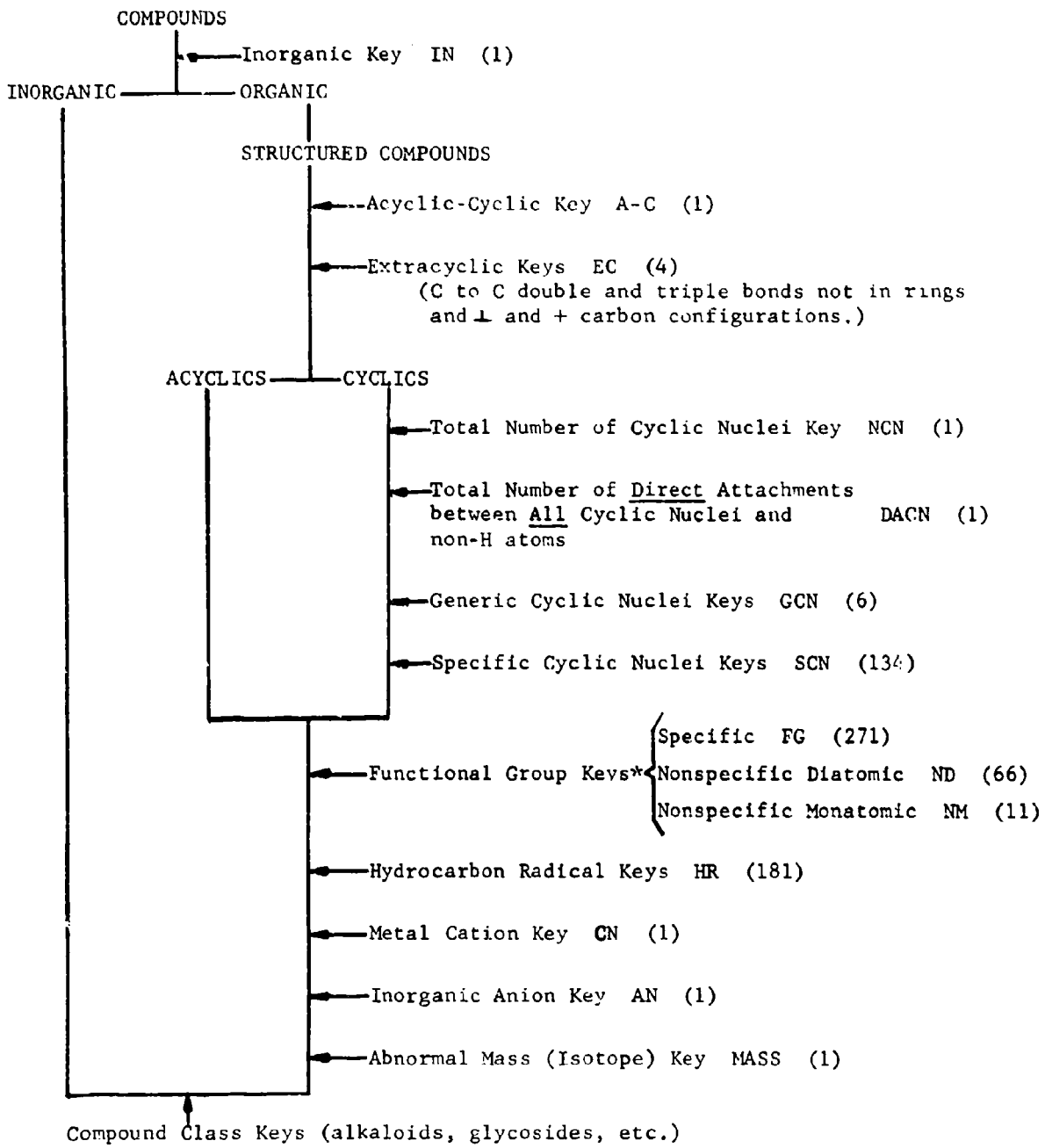
Collectively, the several tasks undertaken during this contract period contribute toward the evolution of a model operational system, suitable for trial and demonstration purposes, from the experimental system developed under Contract DA-18-035-AMC-288(A). More specifically, these tasks are designed to (a) refine the strategies and programs relative to storage, search and output of chemical information and data, (b) adapt the on-line, real-time capability to a multi-terminal system which can accommodate up to four Teletype 35 devices and the Data Products Corp. chemical printer, (c) explore techniques for the effective handling of nonstructural information and data, (d) augment test files suitably to permit demonstrations of structural and nonstructural search and retrieval, and (e) provide the necessary computer search capability for conducting all tests and processing all queries.

It is understood, of course, that the testing and processing mentioned in (e) above is integral to all of the tasks. These operations, together with the required pre-planning and the assay of the findings, account for a large fraction of the Project's total effort. It would serve no useful purpose to provide a detailed account of these day-to-day operations; rather, the ultimate net results find expression only in terms of reported improvements in strategies, programs, techniques, etc.

2. Chemical Search Key Revision

With completion of the assay of the results of large scale exercising of the experimental CIDS (which extended over an 18-month period and involved the processing of hundreds of structural queries), it has been possible to formulate the complete set of molecular and structural search screens which will be incorporated into the model operational CIDS early this fall. Through these keys, provision exists for effective retrieval in terms of qualitative and quantitative molecular composition and substructural characteristics. The structural fragment keys of the experimental system have been modified, deleted, and augmented in accord with the dictates of the experimentation. The complete lexicon of structural keys, which is currently being published in handbook form as the CIDS No. 6 report (1), numbers about 850 individually encoded keys subdivided into several conventional chemical categories. An overview of these categories is presented in Table I. The specific keys which have been admitted for cyclic

TABLE I. CIDS STRUCTURAL SEARCH KEYS OVERVIEW



* Doubly encoded, i.e., attached to a non-ring atom and to a ring atom.

nuclei, functional groups, and hydrocarbon radicals have been selected on the basis of their expected frequency of occurrence in a large unbiased file of compounds. With each class, however, nonspecific or generic keys are provided to encode compounds, or portions of compounds, which are not responsive to the specific keys. As time permits, the lexicon of search keys will be expanded to accommodate certain kinds of compounds which are currently denied admission to the CIDS file, e.g., inorganics, polymers, coordination complexes, etc.

3. Nonstructural Information and Data

It has long been recognized that questions addressed to an operational CIDS will often contain one or more nonstructural parameters. A detailed analysis of 273 "live" questions submitted by 15 different defense installations disclosed that such was the case in upwards of 70 percent of the questions. Most of these questions pertained to the very broad category of information dealing with physical, chemical, and biological properties and applications, although other kinds, such as literature references, sources of supply, methods of synthesis, contractual information, etc., were well represented. It has also been recognized that the total nonstructural area (1) is enormous in scope, (2) cuts across numerous sciences and technologies, (3) includes a variety of nontechnical information of concern to management, and (4) is replete with complications referable to the vagaries of terminology.

One technique for the storage and retrieval of nonstructural material is currently being explored which utilizes an open-ended list of nonstructural descriptors. In its present experimental mode, it employs about 54 such descriptors assembled by Edgewood Arsenal (Table II) with suitable provision for accommodating any number of additional ones as experience discloses their need. In the model operational CIDS, several thousand Army chemical compounds will be tagged with these descriptors in such a way that the documents or literature references containing the detailed information are identified. The technique is described further in the following.

In order to provide the ability to search on the basis of nonstructural information and data (NSID) as quickly as possible, it was decided to incorporate this material into the CIDS system in such a way as to minimize the number of program modifications required. For this reason it was decided to represent the nonstructural information indexes as short mnemonic alphabetic codes which

could be keys in the search system and function like the present CIDS structural keys. Thus queries could be composed of any logical combination of NSID keys and/or structural keys.

It was decided that the NSID would be entered as reference/descriptor sets associated with a pertinent compound. This means that each reference source will provide some set of descriptors selected from the master list (Table II) to be manually assigned to compounds as they are entered in the file.

The NSID codes and related source references (in abbreviated form) are to be accommodated in the nomenclature block of the current CIDS record format. This permits the use of the editing capability of the registry system and the use of current output programs in the search system.

A program has been written under subcontract by the Computer Command and Control Co. to convert the NSID descriptors into search keys after the compound records have been processed through the registry system. This program scans the nomenclature block of a record, extracts all NSID codes from the reference/descriptor sets, and removes any duplicates which may be present (in the event that the same descriptor was associated with more than one reference). Each of these codes is then translated to a CIDS formatted key and stored in the key block of the record. Duplicate keys are not required for retrieval and thus are not generated in order to conserve storage space in the inverted key index. However, each NSID code is printed with its appropriate reference when the compound record is retrieved.

Modifications to the CHEMTYPE system were required in order to permit input of the new information in the typed input record. It was also necessary to introduce a new key type in the search system to permit the use of NSID keys in queries.

TABLE II. NONSTRUCTURAL CATEGORIES

Descriptor	Code
Applications	- AP
Activity Coefficient	- AC
Analytical Detection	- AD
Analytical Determination	- AN
Boiling Point	- BP
Biological Suppressant	- BS
Crystalline Form	- CF
Chromatographic Methods	- CM
Cost	- CO
Critical Pressure	- CP
Color	- CR
Critical Temperature	- CT
Dissociation Constants	- DC
Derivatives	- DV
Entropy	- EN
Electron Spin Resonance Spectrum	- ES
Free Energy	- FE
Geometric Isomers	- GI
Heat Capacity	- HC
Heat of Dilution	- HD
Heat of Formation	- HF
Heat of Solution	- HS
Heat of Sublimation	- HU
Heat of Vaporization	- HV
Hydrates	- HY
Ionization Constants	- IC (pKa, pKb)
Incapacitating Dose (Dosage)	- ID
Infrared Spectrum	- IR
Kinetics of Hydrolysis	- KH
LD_{50} (Dosage)	- LD
(Med) Minimum Effective Dose (Dosage)	- ME

(continued)

TABLE II. NONSTRUCTURAL CATEGORIES (continued)

Descriptor	Code
Melting Point	- MP
Mass Spectrum	- MS
Nuclear Magnetic Resonance Spectrum	- NS
Optical Rotation	- OR
Polarography	- PO
Purification	- PU
Respiratory Inhibition	- RE
Refractive Index	- RI
Raman Spectrum	- RS
Solvent of Crystallization	- SC
Specific Gravity	- SG
Specific Heat	- SH
Hammett Sigma Values	- SI
Solubility	- SO
Specifications	- SP
Surface Tension	- ST
Suppliers	- SU
Solvates	- SV
Synthesis	- SY
Triple Point	- TP
Ultra Violet Spectrum	- UV
Viscosity	- VI
Vapor Pressure	- VP

4. CIDS-Dedicated Computer

The IBM 7040 computer was dedicated to the CIDS project on 1 July 1968. Before that date the Project used the same computer rented by the University Computer Center from IBM and available to all University users. The University Computer Center transferred to an IBM 360-series computer and the Project assumed the 7040 contract with IBM. The 7040 computer is available for CIDS usage (dedicated time) 12 hours per day 7 days per week at a cost of about $14,000 per month. Maximal usage under current costing would thus be much less costly than the same amount of time under the past (rental) costing of $100 per hour. For the period of this report, appreciable savings have been effected on the current basis although computer usage has by no means reached the maximum; the increased usage further contemplated will make this arrangement even more beneficial in terms of comparative costs.

The computer is used extensively in conducting various CIDS R&D operations, both intraproject and between the University facility and Edgewood Arsenal. Major categories of usage include file generation, experimental search, and computer program debugging and testing. A regular program of weekly testing with the Line Printer at Edgewood Arsenal has been instituted. A record disclosing the hourly usage in each category is included in each monthly report to the Project Officer.

5. Improvements in Search Techniques

A major programming effort has been going on since before 1 July on the following:

(1) The system is being modified to use the revised (CIDS No.6) search keys and the corresponding query language.

(2) The size of the data file is being reduced by (a) conversion of the connection table to Mechanical Chemical Code (MCC) whereby the table is compacted to about one-tenth its original size, and (b) compression of the structural formula image (representation of the structural formula by means of coordinates) to about one-half its present size. The MCC was developed at the University under another contract. The notation requires compression and decoding programs and employs about 2.4 characters per nonhydrogen atom.

(3) A new atom-by-atom search program has been written to operate on the connection table produced by the decoding of the MCC. The new program provides increased capability in the specification of structural fragments for atom-by-atom search.

New programs are written in FORTRAN whenever this is possible without detriment to the system, in order that the programming will be usable in the final computer system at Edgewood Arsenal, but no existing satisfactory MAP programs are being rewritten for the sake of having them in FORTRAN.

Detailed documentation on the revised atom-by-atom search program (2) has been received from the Computer Command and Control Co. where the program was written under subcontract. However, the transmission of this report is being delayed until documentation has been written for the subexecutive programs which monitor the output of the atom-by-atom search. A more comprehensive report will be made at that time.

6. Remote Terminal Querying

The multi-terminal real-time retrieval system (CIDS II) has been in operation since the fall of 1968. In the single-terminal on-line system (CIDS I), communication with the search system was possible only through a single terminal at any one time. The system now provides the capability for dialing in queries from a total of four terminals concurrently. Output from queries can be routed either to the teletype asking the question or to the Data Products Chemical Line Printer located at Edgewood Arsenal.

The multi-terminal system was programmed as a special purpose time-sharing system. The design and implementation of the monitor, retrieval scheduler, and terminal input/output facility is presented in a doctoral dissertation (3) by Mr. Paul R. Weinberg. A document by Bonnie Sherr (4) has been prepared as a user's guide for this system. It describes the teletype command language necessary to manipulate the text of a query and run it through to completion.

The cathode ray tube, driven by the DEC-338 computer, will provide another type of communication in which the structure portion of a query would be input by "drawing" it on the cathode ray tube. When the cathode ray tube is in use,

only three other terminals can be used at the same time. Output for it will appear on the cathode ray tube. The Chemical Line Printer will not be available at the same time as the cathode ray tube since they both require the 201 Data Set.

7. Cathode Ray Tube Input/Output

Programming for query input via the CRT has been completed and debugged and is now in the process of system-integration and testing. The program permits construction of, and interprets as chemical structures, diagrams of molecules or molecule-fragments using a light pen and a cathode ray tube. This program will be integrated into the retrieval system in the near future. The user constructs an arbitrary configuration of element-atoms connected by bonds as desired. Each bond is designated as either (a) single, double, or triple, and (b) acyclic, resonant ring, or ring but not resonant. Changes are easily made. On command the 338 computer stores the diagram, and can return it to the tube face after the tube face has been used for other "drawings". The computer also interprets the drawing in a form equivalent t a connection table.

Programming for CRT output has been completed and will be tested with live data from the search system as soon as integration of the CRT input has been achieved.

An interim report (5) describing the results to date on the cathode ray tube as an input-output device in CIDS has been prepared and transmitted to the sponsor. This report will be updated as the work continues and will ultimately issue as a formal CIDS document.

8. File-building

A Digi-data Corp. device for converting from paper tape to magnetic tape has been installed. This converter creates magnetic tape images of the paper tapes produced by the chemical typewriters, and has been in operation since July 1968.

9. File Status

The initial files of compounds in the model operational CIDS will contain between 35,000 and 40,000 compounds. Of these, about 7,000 are from the EA

Toxicological Information Center (TOXINFO) and about 26,300 from the Chemical-Biological Coordination Center (CBCC). The remainder are from a file currently under construction at Edgewood Arsenal and known as the Task 07 File. The information on the TOXINFO and CBCC compounds is now stored on magnetic disk for search in the real time mode and the Task 07 compounds will be handled similarly. All compounds will be amenable to structural search, using the keys of the experimental system until such time as the CIDS No. 6 keys are operable. The Task 07 compounds will be searchable also in terms of the categories of nonstructural information listed in Table II of this report.

A concordance has been prepared which relates the CIDS Registry Number to the Local Control Number(s) and vice versa for each compound in the presently existing 33,300 compound CIDS on-line search file. These listings include the molecular formula(s) of each compound.

Approximately 50,000 additional compounds are being held on paper tape for processing after the MCC notation and the new search keys have been incorporated into the system. Some of these are CBCC compounds and others originated from a variety of files selected by Edgewood Arsenal. Conversion of these records to magnetic tape will begin in the near future.

10. The CIDS No. 6 Report

This formal CIDS publication (1) constitutes the Final Report for Contract DA19-035-AMC-288(A) and documents all chemical search components appropriate to compounds currently admitted to the revised system. The search components subdivide into two general types depending on whether they describe characteristics discernible through computer probes of molecular formulas or structural formulas.

The document is designed to function as a desk-top tool in the intellectual assignment of CIDS chemical search screens to queries addressed to the system. The information is presented from the point of view of a chemist, i.e., it permits stipulation, in conventional chemical fashion, of all features of chemistry appropriate to a query but does not prescribe for the transformation of this information into a formal computer query.

The concluding section of the report provides 52 illustrations of the total assignment of the chemical search keys to a wide structural spectrum of compounds.

11. The CIDS No. 7 Report

This report is visualized as the next in the series of formal CIDS documents and consists essentially of an updating and expansion of an earlier interim report (6) describing the CIDS retrieval language. The updating involves the additions and alterations to accommodate the revised atom-by-atom search program, which was completed some time ago, and the revised lexicon of molecular and structural search screens reported in the CIDS No. 6 document. The expansion consists of (1) additional details of search strategy and (2) a new section displaying an assortment of about 20 user-type structural questions and showing for each (a) the chemical analysis of the question, (b) the assignment of search screens, and (c) the formulation of the fully encoded query. The report is currently in an early stage of initial draft.

LITERATURE CITED

1. Clarence T. Van Meter, Eric N. Goldschmidt, Margaret Milne, Handbook of CIDS Chemical Search Components, CIDS No. 6, University of Pennsylvania, Philadelphia, Pa., December 1968

2. Donald Headley, Documentation: Atom by Atom Search Programs, Computer Command and Control Company, Philadelphia, Pa., December 23, 1968

3. Paul R. Weinberg, A Time Sharing Chemical Information Retrieval System, Ph. D. dissertation, University of Pennsylvania, Philadelphia, Pa., 1969

4. Bonnie Sherr, The CIDS Multi-Terminal Command Language for Teletypes, Project CIDS, University of Pennsylvania, Philadelphia, Pa., October 1968

5. Andre M. Gagnoud, Input-Output of Chemical Structural Formulas, Report #1, University of Pennsylvania, Philadelphia, Pa. November 1968

6. Paul R. Weinberg, The CIDS Retrieval Language, University of Pennsylvania, Philadelphia, Pa., November 1967

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
University of Pennsylvania
Philadelphia, Pennsylvania 19104

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP
NA

3. REPORT TITLE
A CHEMICAL INFORMATION AND DATA SYSTEM

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Semi-annual Report, 1 October 1968 - 28 February 1969

5. AUTHOR(S) (First name, middle initial, last name)
Van Meter, Clarence T., Powers, Ruth V., Plotkin, Morris, and Hill, Helen N.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
28 February 1969	25	6

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
DAAA15-69-C-0140	
b. PROJECT NO.	
c. Task: 2P062101A72702	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
Army chemical and information data systems	Edgewood Arsenal Technical Support Directorate, Edgewood Arsenal, Maryland 21010 (Stanley Goldberg, Proj. O., Ext 5207)

13. ABSTRACT

This document describes research and development activities in progress under Project CIDS at the University of Pennsylvania designed to advance the experimental chemical information and data system to the status of a model operational system. It treats primarily of the revision of the chemical search screens, an exploratory technique for incorporating nonstructural information and data, and various improvements in search techniques. It reports the development and operability of a multi-terminal real-time system which is capable of processing queries concurrently from four terminals, and it summarizes experience with the cathode ray tube as an input-output device. It also reports the status of initial file construction for the model operational system, and the completion and distribution of the CIDS No. 6 Report. The latter is a handbook of all CIDS chemical search components.

DD FORM 1 NOV 1473 REPLACES DD FORM 1473, 1 JAN 64, WHICH IS OBSOLETE FOR ARMY USE.

UNCLASSIFIED

Security Classification

14.

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Chemical information and data system						
CIDS chemical search keys						
Nonstructural information and data						
Nonstructural descriptors						
CIDS computer usage						
Connection table compaction						
Mechanical Chemical Code						
Atom-by-atom search						
Multi-terminal real-time system						
Remote terminal querying						
Cathode ray tube						
Compound files						
CIDS chemical search key handbook						