NBS TECHNICAL NOTE 781

A Study of Six University-Based Information Systems

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards[1] was established by an act of Congress March 3, 1901. The Bureau's overall goal is to strengthen and advance the Nation's science and technology and facilitate their effective application for public benefit. To this end, the Bureau conducts research and provides: (1) a basis for the Nation's physical measurement system, (2) scientific and technological services for industry and government, (3) a technical basis for equity in trade, and (4) technical services to promote public safety. The Bureau consists of the Institute for Basic Standards, the Institute for Materials Research, the Institute for Applied Technology, the Institute for Computer Sciences and Technology, and the Office for Information Programs.

THE INSTITUTE FOR BASIC STANDARDS provides the central basis within the United States of a complete and consistent system of physical measurement; coordinates that system with measurement systems of other nations; and furnishes essential services leading to accurate and uniform physical measurements throughout the Nation's scientific community, industry, and commerce. The Institute consists of a Center for Radiation Research, an Office of Measurement Services and the following divisions:

Applied Mathematics — Electricity — Mechanics — Heat — Optical Physics — Nuclear Sciences[2] — Applied Radiation[2] — Quantum Electronics[3] — Electromagnetics[3] — Time and Frequency[3] — Laboratory Astrophysics[3] — Cryogenics[3].

THE INSTITUTE FOR MATERIALS RESEARCH conducts materials research leading to improved methods of measurement, standards, and data on the properties of well-characterized materials needed by industry, commerce, educational institutions, and Government; provides advisory and research services to other Government agencies; and develops, produces, and distributes standard reference materials. The Institute consists of the Office of Standard Reference Materials and the following divisions:

Analytical Chemistry — Polymers — Metallurgy — Inorganic Materials — Reactor Radiation — Physical Chemistry.

THE INSTITUTE FOR APPLIED TECHNOLOGY provides technical services to promote the use of available technology and to facilitate technological innovation in industry and Government; cooperates with public and private organizations leading to the development of technological standards (including mandatory safety standards), codes and methods of test; and provides technical advice and services to Government agencies upon request. The Institute consists of a Center for Building Technology and the following divisions and offices:

Engineering and Product Standards — Weights and Measures — Invention and Innovation — Product Evaluation Technology — Electronic Technology — Technical Analysis — Measurement Engineering — Structures, Materials, and Life Safety[4] — Building Environment[4] — Technical Evaluation and Application[4] — Fire Technology.

THE INSTITUTE FOR COMPUTER SCIENCES AND TECHNOLOGY conducts research and provides technical services designed to aid Government agencies in improving cost effectiveness in the conduct of their programs through the selection, acquisition, and effective utilization of automatic data processing equipment; and serves as the principal focus within the executive branch for the development of Federal standards for automatic data processing equipment, techniques, and computer languages. The Center consists of the following offices and divisions:

Information Processing Standards — Computer Information — Computer Services — Systems Development — Information Processing Technology.

THE OFFICE FOR INFORMATION PROGRAMS promotes optimum dissemination and accessibility of scientific information generated within NBS and other agencies of the Federal Government; promotes the development of the National Standard Reference Data System and a system of information analysis centers dealing with the broader aspects of the National Measurement System; provides appropriate services to ensure that the NBS staff has optimum accessibility to the scientific information of the world. The Office consists of the following organizational units:

Office of Standard Reference Data — Office of Technical Information and Publications — Library — Office of International Relations.

---

[1] Headquarters and Laboratories at Gaithersburg, Maryland, unless otherwise noted; mailing address Washington, D.C. 20234.
[2] Part of the Center for Radiation Research.
[3] Located at Boulder, Colorado 80302.
[4] Part of the Center for Building Technology.

A Study of Six University-Based Information Systems

Beatrice Marron, Elizabeth Fong,
Dennis W. Fife, and Kirk Rankin

Institute for Computer Sciences and Technology
National Bureau of Standards
Washington, D.C. 20234

Sponsored by

National Science Foundation
18th and G Street, N.W.
Washington, D.C. 20550

U.S. DEPARTMENT OF COMMERCE, Frederick B. Dent, *Secretary*
NATIONAL BUREAU OF STANDARDS, Richard W. Roberts, *Director*

Issued June 1973

National Bureau of Standards Technical Note 781

Nat. Bur. Stand. (U.S.), Tech. Note 781, 98 pages (June 1973)

CODEN: NBTNAE

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402
(Order by SD Catalog No. C13.46:781). Price $1.25 domestic postpaid or $1.00 G.P.O. Bookstore.

CONTENTS

A STUDY OF SIX

UNIVERSITY-BASED INFORMATION SYSTEMS

B. Marron, E. Fong, D. W. Fife and K. Rankin

A methodology for categorically describing computer-based information systems was developed and applied to six university-based, NSF-supported systems. The systems under study all operate as retail information centers primarily serving campus communities by accessing large commercially-available data bases using 3rd generation computer configurations. The systems vary in design philosophy, mode of user service, transferability characteristics, and operational status. A summary matrix is included.

Key words: Computer-based systems; information systems, university; university computer systems.

UNIVERSITY-BASED, NSF-FUNDED, INFORMATION SYSTEMS

PART I - GENERAL

A. INTRODUCTION:

In December of 1971 an interagency agreement was established between the National Bureau of Standards and the National Science Foundation, Office of Science Information Services, whereby the NBS Center for Computer Sciences and Technology was charged to "assess the present status of automated scientific and technological information systems".[1] A number of tasks were defined and studies undertaken to help NSF meet its immediate decision-making needs and to develop methodologies and tools for evaluating and comparing science information activities. This report documents such a study of six university-based, NSF-supported, information systems.

The NSF university-centered effort involves funding to 6 universities: UCLA, U. of Georgia, Lehigh U., Ohio State U., U. of Pittsburgh, and Stanford U. The immediate objectives of NSF support for university-centered information systems are three-fold: (1) to meet the information requirements of academic scientists and the students they are training; (2) to establish "retail" campus-based terminals to accept the "wholesale" machine-readable tapes from society-based, discipline-oriented systems, as well as mission and problem-oriented products from Federal and private sources; and (3) to support the development of major nodes for the emerging national science information system.

This report is organized in two parts.

Part I consists of:

A. this introduction,

B. detail of the methodologies employed,

C. a summary matrix of the systems studied,

---

1. Interagency Agreement #NSF-CA68

D. observations about the university-based systems, and

E. outlook, including conclusions, state-of-the-art assessment, and recommendations.

Part II consists of 6 separate, free-standing reports detailing the 6 systems in a common descriptive format.

B. METHODOLOGY

A few generalizations about NSF involvement in these university systems should be noted. Many of the systems were already underway before NSF funding was secured. NSF support was specifically stated as being for development rather than operation of these systems. Funding extends for 2 to 5 years only, and is reviewable every year. The systems were pledged to operate beyond the development period on their own funds. Most of the systems are centered in the university libraries. The tapes purchased on NSF funds must be science-oriented. Only initial tapes can be purchased with NSF monies. The goals of the various systems are not the same, and the services offered vary in such respects as batch or interactive, retrospective or current awareness service.

One day site vists were made by NBS personnel to each project. As information was gathered, read, and assimilated, a categorical descriptive format was developed for organizing the details on the specific information systems under study. Although developed for the 6 university-based, NSF-supported systems, it is expected that the format is general enough to be applicable to other systems being analyzed. Appendix 1 shows the organization of the categorical description with scope notes. Appendix 2 is a list of data base suppliers, some of which provide input to more than one of the university systems.

Major uses of this study are listed below:

1. General descriptive summary of existing systems in a common format.

2. Comparison of services and systems, primarily on a macro-level.

3. Identification of user-oriented functions and capabilities.

4. Assessment of the potential network relationships of various systems.

5. Evaluation of proposed new systems relative to the state-of-the-art.

6. Identification of transferable programs and general methodology.

7. Highlighting of economic factors in selecting sources for information service.

A draft of this report was sent to the principal investigators at the six universities for review and validation. We gratefully acknowledge their cooperation and the constructive tone of their comments. Where possible we have included their suggestions as well as updates to the basic information which was collected in the summer of 1972.

C. SUMMARY MATRIX:

To assist in an overview of the six systems under study, the summary matrix on page 4 was prepared. More detail can be found in the system reports in Part II.

D. OBSERVATIONS ON THE 6 UNIVERSITY-BASED SYSTEMS:

Some common features are:

1. All 6 operate as retail information centers, providing service to individuals.

2. All primarily serve their campus communities, although they either do, or plan to, serve others on a cost-recovery basis.

3. Five of the 6 offer current awareness services; all 6 offer or plan to offer retrospective search services in the near future, either batch or on-line.

4. All use several large commercially available data bases (although Stanford used only one, MARC).

5. All use large 3rd generation computer systems.

C. SUMMARY MATRIX

University and Project	Hardware	Anticipated Principal Mode of User Service	No. of Commercially Available Data Bases	Names of Commercially Available Data Bases	Operational Status: Current Awareness	Operational Status: Retrospective	Library Relationship	Software Transferability Potential	Special Features
UCLA Center for Information Services	IBM 360/91	Batch and on-line searches and current awareness.	6	BA CA-C CAIN CIJE COMPENDEX RIE	Spring '71 (interim software)	Planned for Fall '72 (interim software)	Joint participation between library and computer center	Moderate, for interim software (IBM TEXTPAC).	Systematic library acquisition & cataloging procedure for machine-readable data bases. Modular software structure in system under development
U. of Georgia Information Center	IBM 360/65	Batch search of large data bases.	18	BA, BIORI, CA-C, CAIN, CBAC, CIJE, CITE, CT, COMPENDEX, GEOL, GEO-REF, GPA, NSA, RIE, SOCAB, SPIN, TOXI, USGRDR	Since May '68	Since May '68	Computer center project--no formal ties to library	Moderate, transferable within 360 family. Documentation & programs available for purchase.	On-line profile entry and editing. Very large production-oriented operation. Many large data bases.
Lehigh U. Project Leadermart	CDC 6400	On-line interactive access to major as well as local data bases.	4	ASCE CA-C COMPENDEX MARC	Since '71	Since '71	Project in the library	Moderate, programmed in FORTRAN. Modules have been run on 360/65 with only I/O statements changed.	Natural language-oriented. Interactive searching with very large data bases.
Ohio State U. Mechanized Information Center	IBM 370/145	Batch current awareness of large data bases.	5	PANDEX, ISI, NTIS, MARC * merged CT	Since Oct. '71	Planned Fall '72	Project in the library	Moderate, programs are not generalized but written in COBOL & PL/1	Multi-disciplinary data base. Used existing software, hence short implementation time. Data base overlap studies conducted.
U. of Pittsburgh Campus Based Inf. System	PDP-10	Batch search of large data bases	6	CA-C CT COMPENDEX GRA NASA ASM/IM	Summer '72	Summer '72	Integrated project among library, computer and academic departments	Low, machine-dependent. Only transferable to PDP-10 with similar peripherals	On-line profile negotiation using subset of the data base. Discipline-oriented centers providing "one-stop" information service. Emphasis on information networks and access to special subscription services.
Stanford U. SPIRES II	IBM 360/67	On-line interactive access to both large and small local data bases.	1	MARC	Not applicable	Planned Fall '72	Computer center project--no formal ties to library	Low, embedded in ORVYL Monitor and written in PL 360	Capable of using WYLBUR text editor. Tutorial features in conversational mode. Generalized re-entrant parser and semantic routines.

6. All, except Stanford, use scientific data bases. Most started with one of the Chemical Abstract Services data bases.

7. Four systems are operational. The other 2 have pilot systems operational.

Some special features worthy of note are:

1. Multidiciplinary data base. Ohio State University has merged 4 large data bases into one and claims great efficiency.

2. On-Line Profile Negotiation. The University of Pittsburgh offers on-line profile negotiation with immediate on-line search of sample data bases for feedback in redefining the profile. The University of Georgia, on the other hand, at present, offers only on-line profile entry and editing without associated search feedback.

3. Common Internal Format. Most of the systems convert all their input data bases to some common internal format so that one set of processing routines services all data bases regardless of contents. Compatibility or convertibility between these formats would enhance networking capabilities.

4. Advanced System Implementation Techniques. The Stanford SPIRES II system employs a syntax parsing technique together with a list of semantic processes. A generalized parser recognizes the user's input commands and executes the commands by executing the appropriate semantic routines. Also the parser and the semantic lists are reentrant and therefore may be readily shared by a number of users. UCLA's system under development employs a modified version of this methodology. These advanced system implementation techniques permit modular top-down programming.

5. Natural Language Processors. Lehigh's system is natural-language oriented. The user states his query in natural English prose, the logical and syntactic structure of which is analyzed by the system. The system responds with a list of its phrases which are the "closest" to the user's query. The user is thus led on to reformulate his query (if he is not satisfied with the

system's response) or else is ultimately presented with a list of relevant documents (if he is satisfied).

6. Information Specialists. Advances in technology have not diminished the role of the information specialist. Experiences at the Universities of Pittsburgh and Georgia point up the need for specialist interface between user and system. The new technology provides better tools for the information specialist, rather than replacements for him.

E. OUTLOOK:

Although this report is primarily descriptive and not intended as a comparative evaluation of the six systems, several broad conclusions are evident regarding the state-of-the-art and the overall accomplishment reflected in these systems.

- An apparent benefit of NSF's funding the development of these university-based systems has been the advancement of the state-of-the-art in the implementation of information systems and services, and in advanced technology such as syntax analyzers, automated profile definition, automatic indexing procedures and thesaurus building.

- Batch-processing system technology (hardware and software) is demonstrably proficient and adequately developed for providing various information services effectively and economically. On-line, interactive technology is basically established for necessary functions, but rather less well developed in production effectiveness for accommodating a large and growing user community.

- Since services are provided without direct cost to university users of these systems, it is difficult to project their economic viability in a competitive market. However, current progress in developing a base of outside paying customers is promising for continued operational evolution of these centers.

- Two contrasting philosophies have arisen regarding the user interface: one calls for the user to interact with the system directly; the other calls for an information specialist as a buffer between the system and the user. Among the six universities studied, Stanford and Lehigh are developing systems in line with the first philosophy, while Ohio State, Georgia, UCLA and Pittsburgh emphasize more the latter. There is no obvious or simple choice

on this issue, but one can appreciate that a high production environment with a variety of heavily used data bases puts considerable stress on user training and assistance through information specialists.

- The technical means of transferring a program or a data base are still inadequate and irregular. A solution for achieving greater program transferability involves the use of higher-level languages that are standardized. However, the input/output operations that are heavily involved in information system programs are primarily accomplished by the resident operating system, which introduces machine dependence. Data base transferability may be enhanced by storing the data definitions along with the data values. However, the characteristics of mass storage devices impose various data arrangements, which limit the effectiveness of this technique as well.

- The close relationships with campus libraries that are emphasized in most of these systems have established automated information services as a highly desirable, perhaps necessary augmentation of traditional library service. For economy and responsive service, it follows that each university should have a minimal automated information system tailored to the needs of its user community. Such replication of capability among campuses is not redundant just as the multiplicity of libraries is not redundant. Techniques need to be worked out, however, for the equivalent of an inter-library loan system, i.e., access to data bases and/or services not available on a home campus. Further study of nationwide cooperative information networking is therefore indicated.

- Serious criticisms have been raised about the quality of the data files received by the University centers from wholesale sources. For example, it was stated by one center that editing and correcting some data bases costs as much as the original purchase of the data bases. Data quality control clearly needs a deeper analysis and perhaps further development at the wholesale level.

Appendix 1

CATEGORICAL DESCRIPTIVE FORMAT

1. GENERAL DESCRIPTION

   1.1 Project and Organization Name:
   1.2 Brief Statement of Operational Philosophy, Objectives, System Description, and Services Offered:
   1.3 Historical Background:
   1.4 Present Operational Status:
   1.5 Future Plans:

2. ADMINISTRATIVE DETAILS

   2.1 Principal Investigator:
   2.2 User Community:
   2.3 Documentation:
   2.4 Services and Cost:

3. DATA BASES

   3.1 Data Base Selection
      3.1.1 Requirements and Acquisition Strategy:
      3.1.2 Discipline Coverage:
   3.2 Data Preparation, Conversion, and Entry (indexing procedures, data element definition,...):
   3.3 Data Base Contents (See also Appendix 2 - DATA BASE SUPPLIERS)
      3.3.1* Data Base (1)* (Brief description including frequency of issue, size, growth rate, cost,...):

4. HARDWARE CONFIGURATION

   4.1 Main Frame:
   4.2 Core Size:
   4.3 Mass Storage Devices:
   4.4 Input Devices:
   4.5 Output Devices:

---

* Repeat this subsection for each data base.

5. SOFTWARE CONFIGURATION

5.1 Operating System:
5.2 Operational Environment (dedicated, partitioned, job mix,...):
5.3 Information System
    5.3.1 Name and Brief Description:
    5.3.2 Source Language:
    5.3.3 Mode (for search, maintenance, profile definition):
    5.3.4 Generalized Packages Used:
    5.3.5 Availability:

6. COMPUTER PROCESSING FUNCTIONS

6.1 Data Definition:
6.2 Data Maintenance:
6.3 Data Retrieval:
6.4 Data Output:
6.5 Special (Input editing and validation, Back-up, Restart and Recovery, Data Security, Tutorial, Browsing,...):

7. USER INTERFACE

7.1 Current Awareness Service Available?:
    7.1.1 Profile Definition:
    7.1.2 Output Form:
    7.1.3 Frequency:
    7.1.4 Charges:
7.2 Batch Retrospective Search Available?:
    7.2.1 Query Formation:
    7.2.2 Response Time:
    7.2.3 Charges:
7.3 Interactive Retrospective Search Available?:
    7.3.1 Query Language:
    7.3.2 Training and/or Assistance:
    7.3.3 Response Time:
    7.3.4 Charges:
7.4 Document Delivery Service Available?:
    7.4.1 Form of Delivery:
    7.4.2 Charges:

8. TRANSFERABILITY CHARACTERISTICS AND EXPERIENCES

8.1 Software Transferability:
8.2 Data Base Transferability:
8.3 Methodology Transferability:

Appendix 2

DATA BASE SUPPLIERS

1. American Geological Institute
   GEO-REF

2. American Institute of Physics
   Searchable Physics Information (SPIN)

3. American Society of Civil Engineers
   ASCE journal abstracts

4. American Society for Metals
   Metal Abstracts Index on Magnetic Tape (METADEX)

5. BioSciences Information Service
   Biological Abstract (BA)
   Toxitapes (TOXI)
   Biosearch Index (BIORI)

6. Chemical Abstract Service
   Chemical Title (CT)
   CA - Condensates - odd and even issues (CA-C)
   Chemical - Biological Activities (CBAC)

7. CCM Information Sciences, Inc.
   PANDEX Current Index to Scientific and Technical Literature (PANDEX)

8. Engineering Index, INC.,
   Computerized Engineering Index (COMPENDEX)
   Current Information Tapes for Engineers (CITE)

9. ERIC - Office of Education
   Research in Education (RIE)
   Current Index to Journals in Education (CIJE)

10. Institute for Scientific Information
    Institute for Scientific Information source tapes (ISI)

11. Library of Congress
    MARC tapes (MARC)

12. National Agricultural Library
    Cataloging and Indexing System (CAIN)

CATEGORICAL DESCRIPTION OF UCLA INFORMATION SYSTEM

Prepared for: Office of Science Information Services
National Science Foundation

Prepared by: Systems Development Division
Institute for Computer Sciences and Technology
National Bureau of Standards

January 1973

CONTENTS

CATEGORICAL DESCRIPTION OF UCLA INFORMATION SYSTEM

1. GENERAL DESCRIPTION

1.1 Project and Organization Name:

Center for Information Services (CIS), UCLA Campus Computing Network, University Library and Institute of Library Research

1.2 Objectives and Operational Philosophy:

The objectives of CIS were stated on page 1 of their "Phase III - Continuation of A Proposal for Development of a Center for Information Services" and are restated here as follows:

(1) CIS should be operational -- designed to meet the daily needs of the University community, and not simply be a research system or experiment;

(2) It should be a general-purpose system -- able to accept a wide variety of both existing and future data and able to satisfy a wide variety of requests;

(3) It should be adaptable -- able to meet unanticipated needs and able to provide feedback on its operation;

(4) It should be replicative -- designed to be able to be installed in many places;

(5) It should be easy to use -- designed to encourage receptivity and use;

(6) It should be an extension of the Library -- designed so that library personnel can integrate it into their operation;

(7) It should be designed as a potential node in an inter-university network.

1.3 Historical Background:

The project started in July 1966. The development plan is defined as follows:

Phase	Dates	Title
I	July 1966-Dec. 1967	Feasibility and preliminary specifications
IIA	July 1969-Dec. 1970	Requirement specification and basic design
IIB	Jan. 1971-June 1972	Detailed design and prototype development
III	July 1972-June 1973	Prototype operation and system completion
IV	July 1973-June 1974	Implementation and evaluation of services

Phase I exploratory work has been done by the Institute of Library Research. NSF/OSIS awarded Phase IIA of this grant in July 1969 to initiate development of a "Center for Information Services" on the UCLA Campus.

Phase IIA goals were to develop specifications for the Center, programs for the use of multiple data bases, and experimental experience with mechanized information services. Results are documented in the seven-part Phase IIA Final Report. Phase IIA was performed under the auspices of the Institute of Library Research with the joint participation of the Institute, the Campus Computing Network (CCN), and the University Library.

Phase IIB responsibility was awarded to the Campus Computing Network, as planned, when task emphasis changed to detailed program specifications and design, prototype development, and the offering of experimental data base services using interim software. Joint participation continues with the Institute principally concerned with coordination, documentation, reporting, test and evaluation, and the user interface. Basic Library concerns include service aspects, development of procedures, and acquisition of data bases. After Phase III installation and shakedown, project responsibility will then be transferred to the University Library

for system operation, evaluation, and maintenance, with CCN and Institute support.

1.4 Present Operational Status:
As of January 1973, CIS was in Phase III, prototype operation and system completion.

1.5 Future Plans:
Future plans consist of the implementation of Phase III and Phase IV. For Phase III tasks, the Library and CCN will jointly operate the prototype Information Processing System (IPS) while the remainder of the software system is completed, and procedures for the fully operational CIS are developed. Phase IV begins on July 1, 1973 with the formal transfer of responsibility for the CIS project from CCN to the Library. The major task for this phase will be to accomplish the transition from a developmental project to a smooth-running operational service. Additional software services for users will be added based upon users' experiences and requests, and the IPS will be tuned as needed to insure an efficient, cost effective operation.

2. ADMINISTRATIVE DETAILS

2.1 Principal Investigator:
William B. Kehl, Director
Campus Computing Network
UCLA, Los Angeles, California
(213) 825-7511

2.2 User Community:
The university community.

2.3 Documentation:

(1) Phase IIA Final Report. Center for Information Services Detailed Systems Design and Programming. NSF Grant GN-827. March 1, 1971. Institute of Library Research, UCLA.

(2) Watson, P. G., and Briggs, R. B., "Computerized Information Services for the University Community." In: Information Storage and Retrieval, Vol. 8, pp. 21-33. Pergamon Press 1972.

(3) Phase III, Continuation of a Proposal for Development of a Center for Information Services, a joint project of the: Campus Computing Network, University Library and Institute of Library Research, UCLA.

(4) Guide for Preparing TEXT-PAC Profiles, Center for Information Services; a joint project of the Campus Computing Network, University Library, and Institute of Library Research, UCLA.

(5) Supplemental Guide: Preparing TEXT-PAC Profiles for the CA-Condensates File, Center for Information Services, a joint project of the Campus Computing Network, University Library, and Institute of Library Research, UCLA.

(6) Annual Report, Institute of Library Research UCLA, July 1970 - June 1971.

2.4 Services and Costs:
At present, CIS is just providing current awareness services, but retrospective searches will be available by the fall of 1972. Now services are free, but plans call for charging non-University of California subscribers, beginning early in 1973.

3. DATA BASES

3.1 Data Base Selection:

3.1.1 Requirement and Acquisition Strategy:

The data base acquisition problem was carefully considered during Phase IIA. A number of user committees were established not only to provide information, guidance, and support, but also to participate actively in data base selection and experimental utilization. The user committees are discussed on page 24, Part 4 of Phase IIA, Final Report, "Development Scheduling and Planning" by Robert L. Carmichael, dated April, 1971.

The requirements were not only to provide current awareness service to mechanized data bases, but also to be capable of processing small private data bases. The data is not restricted to bibliographies but includes numerical data such as census tapes and full text files.

3.1.2 Discipline Coverage:
The primary data bases cover scientific and technical disciplines, and include CAIN, COMPENDEX, CA-Condensates, ERIC files, and Biological Abstracts. They are considering acquisition of APA, INSPEC, and maybe MARC and Food Service and Technology Abstracts.

3.2 Data Preparation, Conversion and Entry:
Data is processed in the format in which it is received; it is not converted to a common format. Extract routines resolve incompatibilities between file data formats and program data formats at data access time.

3.3 Data Base Contents:
The following are the data bases available through the current awareness services using interim software (IBM's TEXT-PAC).

3.3.1 Data Base (1): CA-Condensates (odd and even issues). Ca-Condensates is the computer searchable complement to the printed publication, Chemical Abstracts (CA), which covers the full range of chemistry, referencing 250,000 articles per year.

CA-Condensates is issued weekly; the content corresponds to the issue of CA. The tape version, CA-Condensates, precedes the corresponding printed issue of CA by several weeks due to the time required to print, bind, and distribute CA printed issues.

The abstracts in CA and CA-Condensates are grouped into five categories: Biochemistry, Organic Chemistry, Macro-molecular Chemistry, Applied Chemistry and Chemical Engineering, and Physical and Analytical Chemistry. The first two groupings are published as an odd numbered issue one week, and the last three groupings are published as an even numbered issue the following week.

The UCLA collection started Jan. 1971. There are 123 profiles for the even issues and 204 profiles for the odd issues and currently 149 users are served.

3.3.2 Data Base (2): Cataloging and Indexing System (CAIN). The Cain file, issued

monthly by the National Agricultural Library, contains bibliographic data on journal articles and monographs encompassing the broad field of agriculture, including agricultural economics and rural sociology, agricultural products, animal industry, engineering, entomology, food and human nutrition, forestry, pesticides, plant science, soils and fertilizers, and other related subjects, previously included in the Bibliography of Agriculture, the National Agricultural Library Monthly Catalog, and the Pesticides Documentation Bulletin. All references cited in the American Bibliography of Agricultural Economics, a publication issued by the American Agricultural Economic Association beginning February 1970, are also included.

The UCLA collection started January 1972 when the University of California at Davis subscribed to the file and elected to have the processing done by CIS. There are currently 110 profiles and 99 users.

3.3.3 Data Base (3): Computerized Engineering Index (COMPENDEX). COMPENDEX, issued monthly by Engineering Index, Inc., is the computer-readable version of the printed publication, Engineering Index Monthly, which contains references spanning all engineering disciplines. These references are taken from professional and trade journals, publications of engineering organizations, papers from conferences and symposiums, and books and other documents.

The UCLA collection started January 1972 and service started April 1972. There are currently 43 profiles and 31 users.

3.3.4 Data Base (4): Educational Resources Information Center (ERIC). Educational Resources Information Center (ERIC) is a nation-wide information network for acquiring, selecting, abstracting, indexing, storing, retrieving, and disseminating the most significant and timely educational research reports and projects. It is a project of the U. S. Office of Education, and produces two data bases, RIE (Research in Education) and CIJE (Current Index to

Journals in Education). RIE and CIJE are produced monthly as printed publications and quarterly on magnetic tape.

At UCLA back issues of RIE are available from 1966 and back issues of CIJE are available from 1969. Service started in April 1972, both having 49 profiles and 40 users.

4. HARDWARE CONFIGURATION

4.1 Main Frame: IBM 360/91

4.2 Core Size: 250-300K partition

4.3 Mass Storage Devices - Tapes

4.4 Input Devices: Tapes, card reader, consoles (CRT, TTY, 2741), RJS (Remote Job Service) stations

4.5 Output Devices: Tapes, printers, consoles (CRT, TTY, 2741), RJS Stations.

5. SOFTWARE CONFIGURATION

5.1 Operating System:
IBM OS/360 MVT. (TSO optional)

5.2 Operational Environment:
The UCLA Campus Computing Network's facilities may be accessed through a number of batch and interactive systems which include APL, ARPA Network access, RJS (Remote Job Service), Quickrun (primarily for fast turnaround of student jobs), OLMS (On-line Mathematical System, a version of the Culler-Fried System), TSO (IBM's time-sharing option), and URSA (A conversational, CRT display oriented, job entry system with powerful file and utility operations). The IPS will operate either as a batch process or under TSO which allows it to have interactive capabilities.

5.3 Information System:

5.3.1 Name and Brief Description:
The software system under development is known as the Information Processing System (IPS). It is designed to be a multi-service, multi-data base program having both batch and interactive facilities. It

consists of four principal sets of programs: a monitor, an analyzer, a set of input/output (I/O) routines, and the services that perform the actual work on the data base.

The I/O routines form an interface between the data base and the service programs, thus allowing the service programs to be format independent. A particular service program can operate on any data base provided the appropriate I/O routines have been selected. This approach allows additional format-independent service programs to be freely created and added to the system. A new data base format can be processed by adding appropriate I/O routines to the system.

5.3.2 Source Language:
Primarily in assembly code with a little PL/1

5.3.3 Mode:
The system may be run in either a batch or interactive mode. When operating in the interactive mode, an autobatch facility will automatically queue for batch processing those jobs requiring lengthy computation and/or files not normally kept on-line to the computer.

5.3.4 Generalized Packages Used:
The IPS is itself a generalized information processing system. (The interim system uses IBM's TEXT-PAC).

5.3.5 Availability:
The IPS will be available from UCLA but arrangements have not been worked out yet.

6. COMPUTER PROCESSING FUNCTIONS

6.1 Data definition:
An I/O package is dynamically configured for a given file based on tables describing the service programs' requirements and the format of the selected data sets. External data definition consists of preparing these tables and cataloging them within the system.

6.2 Data Maintenance:
Maintenance for data bases acquired from suppliers consists only of adding new records. For locally developed data bases, maintenance is handled via an IPS service.

6.3 Data Retrieval:
Data retrieval is handled first by the "Analyzer", which analyzes a user request, determining the data paths, and then setting up the parameter lists for the "service" routines and the I/O package generated.

6.4 Data Output:
The formats for data output are generated in the same manner as for input (see Section 6.1 above).

6.5 Special:
(1) Own-Code facility will allow a user to write a PL/1 program to be run under the IPS, e.g. as a subservice.

(2) Autobatch facility will automatically queue for batch processing those jobs requiring lengthy computation and/or files not normally kept on-line to the computer.

(3) File sharing facility will provide the ability for multiple, independent, IPS services to operate on the same largescale sequential data file during a single pass.

(4) IPS will contain a set of accounting and statistics-gathering capabilities.

7. USER INTERFACE

7.1 Current Awareness Service Available? Yes with interim software. (IBM's TEXT-PAC).

7.1.1 Profile Definition:
For the interim current awareness service, reference librarians serve as profile analysts to assist users with preparing profiles. The profiles are then entered and updated via an interactive URSA processor. Syntax checking is performed by a batch program.

For the IPS, interactive profile negotiation with immediate syntax checking will be available. Text searches against an on-

line sample data base are planned for the future.

7.1.2 Output Form:
Under the interim system, only computer printouts are available. Entries are separated by dashed lines. Limited information on the profile term(s) that caused the hit is provided.

For the IPS, users can specify the desired output format and medium, including CRT display or transmission to a remote site via an existing network.

7.1.3 Frequency:
Quarterly, monthly, biweekly, or weekly, depending upon the frequency of issue of the data base.

7.1.4 Charges:
Free to the university users.

7.2 Batch Retrospective Search Available? Yes. The retrospective searches were only run occasionally during Phase IIB and usually only to bring a user up-to-date on his current awareness searches. A regularly scheduled retrospective service will begin early in Phase III.

7.3 Interactive Retrospective Search Available? Planned for the future.

7.4 Document Delivery Service Available? No. An experiment on this aspect is planned for Phase III.

7.4.1 Form of Delivery:
A user will receive, along with the list of citations, a list of document reference numbers. He may then check the numbers of any documents which he wants to see and return the list to CIS; documents will then be delivered to him by project staff.

7.4.2 Charges:
Cost of any Xeroxing.

8. TRANSFERABILITY CHARACTERISTICS AND EXPERIENCE

8.1 Software Transferability:

The interim software is a generalized package developed by IBM called TEXT-PAC. It consists of Assembly language programs and could be transferred with little effort to other IBM machines.

The IPS will be transferable to an IBM S/360 Model 50 or above with at least 500K bytes of storage; or to an IBM S/370 Model 145 or above with virtual memory. The system will operate optimally when run under OS MVT with TSO (or OS/VS2; it will, however, operate batch only, under OS MFT.

8.2 Data Base Transferability:

The search tapes for the interim system are conveniently usable only at another site running TEXT-PAC on IBM equipment.

No restrictions on data base transferability will be present when the IPS is operational because the tapes will remain in the supplier's format.

8.3 Methodology Transferability:

The format-independent approach to process multiple data bases may turn out to be a noteworthy addition to the state-of-the-art.

The open-ended design of the IPS by defining service calls and implementing service routines is a methodology worth noting.

UCLA is experimenting with original document delivery service. They are facing some administrative problems such as copyright. UCLA is also performing an analysis of journals cited in comparison to holdings.

UCLA library has a very systematic library acquisition and cataloging procedure for their machine-readable data files. Detail is described on page 12 of "Phase III - Continuation of a Proposal for Development of a Center for Information Services."

CATEGORICAL DESCRIPTION OF

UNIVERSITY OF GEORGIA INFORMATION CENTER

Prepared for: Office of Science Information Services
National Science Foundation

Prepared by: Systems Development Division
Institute for Computer Sciences and Technology
National Bureau of Standards

January 1973

CONTENTS

CATEGORICAL DESCRIPTION OF

UNIVERSITY OF GEORGIA INFORMATION CENTER

1. GENERAL DESCRIPTION

1.1 Project and Organization Name:
University of Georgia Information Center

1.2 Objectives and Operational Philosophy:
The University of Georgia Information Center is designed to serve primarily the faculty, research staff and graduate students in the Georgia University System, although it also provides services, on a cost-recovery basis, to other academic institutions, government agencies and commercial organizations. It operates within the Computer Center and is considered by the University administration as an extension to the traditional library services normally available. For subject areas in which bibliographic data bases are available in computer-readable form, the Information Center assists University personnel by performing a computer search for bibliographic references pertinent to the researcher's interests.

The system has been designed primarily for large volume search operations. The search segment of the system is batch-oriented but it provides for on-line data entry of search profiles. Both current awareness and retrospective searches are available on all operating data bases.

The Computer Center also provides the personnel and computer resources necessary to support this program in terms of software development of the necessary computer systems. The Information Center is service-oriented toward the provision of search services; applied research and development tasks are pursued only to the extent that such activities directly support the extension of service facilities. There is no formal connection between the Information Center and the University Library; however, there is a close, cooperative working relationship between the staffs of the two organizations. The Library has absorbed the increased workload of supplying original documents which has resulted from the computer search services.

1.3 Historical Background:
Early in 1966 the University of Georgia began a program to introduce computer-based information services for the faculty and staff on the campus. The program was instituted by the Computer Center with the administrative and financial support of the office of the Vice President for Research for the University. Initially, the service was limited to current awareness searches against one data base in the area of biochemistry, and users of the search service were asked to do their profile construction. The system has grown from 20 profiles on one data base in May 1968, through 300 users with 2000 profiles on 14 data bases in August 1970, to some 900 users with 4400 profiles on 18 data bases in June 1972. Retrospective search capabilities have been added and services have been extended beyond the University of Georgia campus to include all 26 schools in the University System of Georgia, as well as other universities and colleges, government agencies, and industries throughout the United States.

1.4 Present Operational Status:
The University of Georgia Information Center is currently operational. As of June 1972, it was serving some 900 users, mostly in the UGA System, with current awareness and retrospective searches on 18 data bases.

1.5 Future Plans:

a) Expansion of the search services and user community of the Information Center.

b) Development and evaluation of an experimental information dissemination network with the University of Pittsburgh.

c) Developmental work on chemical structure data bases.

2. ADMINISTRATIVE DETAILS

2.1 Principal Investigator:
Dr. James Carmon
Director of Computer Center
University of Georgia
Athens, Georgia 30601
(404) 542-3106

2.2 User Community:
The following statistics on UGA users are from the 1970-71 annual report to NSF referenced in Section 2.5 below.

Classification of Center Users by Type

Type of User	FY '71
University System of Georgia	840
Commercial Users	44
Academic Institutions	9
Experimental Users	4
Total:	897

Geographic Location of Users

University System of Georgia	
Athens (UGA)	796
Other Georgia Institutions	44
Commercial	
Northwest	0
West and Southwest	3
Central	3
South and Southeast	21
North and Northeast	14
Europe	2
Academic	
Northwest	0
West and Southwest	4
Central	1
South and Southeast	2
North and Northeast	2

2.3 Documentation:

1. Final Report to NSF on Grant GN-851 to Expand the University of Georgia Information Center, Oct. 1, 1969 - Sept. 30, 1972.

2. Annual Report to NSF on Grant GN-851 to Expand the University of Georgia Information Center, Oct. 1, 1970 - Sept. 30, 1971.

3. Annual Report to NSF on Grant GN-851 to Expand the University of Georgia Information Center, Oct. 1, 1969 - Sept. 30, 1970.

4. UGA Text Search System, System Description and Program Documentation.

5. UGA Text Search System, Data Base Description and Conversion Programs.

6. UGA Text Search System, Operations Manual.

7. Profile coding and management manual for the University of Georgia Text Search System.

2.4 Services and Costs: (Prices cited are for FY-72. The pricing structure for outside use is reviewed annually on a July fiscal year basis.)

Documentation Price - $100.00
Program Price - $4,000

Services free to members of the Georgia University System. Prices to others are:

Current Awareness Service

Price per profile per issue

Data Base	Price
CA-Condensates	
odd and even	$10.
odd or even only	5.
Chemical Titles	5.
Chemical-Biological Activities	5.
Biological Abstracts	
BA	5.
BIORI	5.
Nuclear Science Abstracts	5.
Compendex	5.

Retrospective Search Service

Price per search

Data Base	Volume* Price	Year
CA-Condensates		
odd and even	$ 70.	$140.
odd or even only	35.	70.
BA-Previews		
BA and BIORI	105	105
BA only	70	70

BIORI only	$ 35.	$ 35.
CBAC	35.	70.
CT	70.	70.
Compendex	70.	70.
NSA	70.	70.
RIE (through 1970)		75.
CIJE	35.	35.

* Volumes correspond to the publishers' printed publication volumes.

3. DATA BASES

3.1 Data Base Selection

3.1.1 Requirements and Acquisition Strategy:

New data bases will be added as user groups large enough to justify services to them are identified. In order to make the service useful to the widest segment possible in the University community, the basic guideline for extension has been to increase the number and scope of the available search services to as many subject areas or disciplines as possible.

3.1.2 Discipline Coverage:

The subject fields covered were initially heavily oriented towards biology, chemistry, and biochemistry. Areas added include engineering, physics, geology, medicine, agriculture and education. Some interdisciplinary data bases are also included.

3.2 Data Preparation, Conversion and Entry:

All data bases received by the Center are converted by Computer program to Standard File Format (SFF) for search with the UGA Text Search System. The search programs have been written for the SFF file structure itself and are essentially independent of any specific data contained in the file. All programs use a common set of tables for data element identification numbers and their corresponding characteristics. Any new data base can be added to the system by converting its format to SFF and by adding to the data element tables any new types of data contained in that file. Conversion programs are written for each new data base. An estimated two weeks is required for writing each program. No indexing or abstracting is done.

The data elements are those provided by the data base supplier.

3.3 Data Base Contents:
The total number of records available in SFF and being searched routinely by the Center at the end of FY '71 was over 2.8 million documents. The growth rate of records for currently received data bases (i.e., those used for current awareness) is approximately 750,000 per year.

The size of the retrospective collection by data base as of July 1, 1971 is summarized below. Descriptions of the data bases follow.

Data Base	No. SFF Documents
BA	315,656
BIORI	219,999
CA-C	779,675
CAIN	131,761
CBAC	90,956
CIJE	31,639
CITE	18,162
CT	782,075
COMPENDEX	131,067
GEOL	7,321
NSA	255,892
RIE	44,080
SPIN	22,382
TOXI	6,604
	2,837,269

3.3.1 Data Base (1): Chemical-Biological Activities (CBAC)

CBAC provides up-to-date coverage of published work concerned with the interaction of organic compounds (drugs, pesticides, etc.) with biological systems (man and other plants and animals). Also covered are metabolism studies and studies of in-vitro chemical reactions of biochemical interests. Nearly 700 journals are monitored. CBAC is a consolidated source of information in the chemical-biological field, and especially appeals to scientists involved in biochemical and physiological research.

The tape version of CBAC corresponds to the Chemical-Biological Activities hard-copy

version; however, tape issues are usually available two-four weeks ahead of the printed copy. Each issue contains 500-600 digests.

3.3.2 Data Base (2): Chemical Titles (CT)
CT, which is issued by Chemical Abstracts Service, contains journal references to approximately 4,500 articles per issue appearing in 650 important U. S. and non-U. S. chemical and chemical engineering journals. Titles that appear in CT represent over 65% of the total abstracts that later appear in Chemical Abstracts. Chemical Titles offers journal references to articles approximately 70 days before their abstracts are published in Chemical Abstracts. In many cases titles appear in Chemical Titles before the journal containing the article is published. Thus Chemical Titles is valuable as an alerting service.

3.3.3 Data Base (3): CA-Condensates (CA-C)
CA-C is the computer searchable complement to the printed publication, Chemical Abstracts (CA), which covers the full range of chemistry, referencing 250,000 articles per year.

CA-Condensates is issued weekly; the content corresponds to an issue of CA. The tape version, CA-Condensates, precedes the corresponding printed issue of CA by several weeks due to the time required to print, bind, and distribute CA printed issues.

The abstracts in CA and CA-Condensates are grouped into five categories: Biochemistry, Organic Chemistry, Macro-moleculer Chemistry, Applied Chemistry and Chemical Engineering, and Physical and Analytical Chemistry. The first two groupings are published as an odd numbered issue one week, and the last three groupings are published as an even numbered issue the following week. Searches may be limited to odd or even numbered issues if desired.

3.3.4 Data Base (4): Nuclear Science Abstracts (NSA). NSA, published semi-monthly by the U. S. Atomic Energy Commission, provides international coverage of the literature on nuclear science and technology. Publications include the scientific and technical reports of the USAEC and its contractors, other government agencies, universities, industrial and research organizations as well as patents, books and journal literature on a world-wide basis.

The subject matter coverage of NSA includes nuclear chemistry, and physics, instrumentation, reactor technology, engineering, earth science and life science.

3.3.5 Data Base (5): Bioresearch Index (BIORI) The tape version of Bioresearch Index (BIORI) corresponds in coverage to the printed copy of this BioScience Information Service of Biological Abstracts. BIORI, published monthly with more than 7,000 titles, includes research reports from symposia, congresses, conferences, reports, bulletins, etc. in which biology is the main emphasis. Biological research papers of a more applied nature are covered, also, in BIORI. Coverage is not duplicated in BA. The tape version of BIORI precedes the printed issue of BIORI by several weeks due to the time required to print, bind and distribute BIORI hard copy.

3.3.6 Data Base (6): Biological Abstracts (BA) The tape version of Biological Abstracts (BA) corresponds in coverage to the printed copy of this BioScience Information Service (BIOSIS) - more than 7,000 journals per year producing more than 200,000 abstracts per year. BA issues, containing more than 5,600 titles, are published semi-monthly and chiefly contain abstracts of research papers in biology published in monthly periodicals. Information covered reflects basic research in biology. The tape version of BA precedes the printed issue of BA by several weeks due to the time required to print, bind and distribute BA hard copy.

3.3.7 Data Base (7): Educational Resources Information Center (ERIC). Educational Resources Information Center (ERIC) is a nation-wide information network for acquiring, selecting, abstracting, indexing, storing, retrieving, and disseminating the most significant and timely educational research reports and projects.

The ERIC program has two component parts: Research in Education (RIE), and Current Index to Journals in Education (CIJE).

RIE is a monthly abstract journal announcing recently completed research and research-related projects in the field of education. The magnetic tape data base is issued quarterly.

CIJE is devoted exclusively to periodical literature, providing detailed indexing for articles in over 500 educational journals. CIJE is a companion to RIE. The printed publication is issued monthly and the computer readable data base quarterly.

3.3.8 Data Base (8): Cataloging and Indexing System (CAIN). The CAIN file, issued monthly by the National Agricultural Library, contains bibliographic data encompassing the broad field of agriculture, including agricultural economics and rural sociology, agricultural products, animal industry, engineering, entomology, food and human nutrition, forestry, pesticides, plant science, soils and fertilizers, and other related subjects, previously included in the Bibliography of Agriculture, the National Agricultural Library Monthly Catalog, and the Pesticides Documentation Bulletin. All references cited in the American Bibliography of Agricultural Economics, a publication issued by the American Agricultural Economic Association beginning February 1970, are also included.

3.3.9 Data Base (9): Computerized Engineering Index (COMPENDEX). COMPENDEX, issued monthly by Engineering Index, Inc., is the computer-readable version of the printed publication, Engineering Index Monthly,

which contains references spanning all engineering disciplines. These references are taken from professional and trade journals, publications of engineering organizations, papers from conferences and symposiums, and books and other documents.

3.3.10 Data Base (10): U. S. Gov. R&D Reports (USGRDR). USGRDR tapes cover the government reports issued by the National Technical Information Service. This data base has recently been renamed Government Report Announcements (GRA).

3.3.11 Data Base (11): Searchable Physics Information Notices (SPIN). SPIN tapes are issued by the American Institute of Physics and contains bibliographic citations to the physics literature.

3.3.12 Data Base (12): Toxitapes, (TOXI). The Toxitapes data base is a machine-readable file, containing bibliographic data and index entries for toxicology literature selected in large part but not exclusively from that announced in Biological Abstracts (BA). The items are categorized as industrial and/or pharmaceutical, with approximately 2000 document entries for each of three update tapes, giving a total file size of 6000 documents.

(The Georgia Information Center entered into an agreement with BioSciences Information Service (BIOSIS) in 1971 to participate in the Toxitapes Experimental Project. Under this agreement Georgia installed the Toxitapes for search using the UGA Text Search System and provided a report to BIOSIS on their use.)

3.3.13 Data Base (13): GEO
GEO is a bibliographic file of North American Geology from the U. S. Geological Survey.

3.3.14 Data Base (14): Current Information Tapes for Engineers (CITE). CITE, issed by Engineering Index, Inc., covers applications technology in Plastics and Electrical/ Electronics Engineering from 300 journals.

The tape records include a searchable segment (index terms) and a display segment (citation, title, author). Abstracts are not a part of the tape record, but are available in hard copy and microfiche.

3.3.15 Data Base (15): Geological Reference File (GEO-REF). GEO-REF, issued by the American Geological Institute, covers the earth sciences, including marine geology, geochemistry, geochronology, geohydrology, geomorphology, etc. There are 1,300 journals reviewed and an average of 7 index terms assigned each entry. The records include author, title, journal title, publication dates, UDC number, and index terms. Tapes are issued monthly.

3.3.16 Data Base (16): Geophysical Abstracts (GPA) GPA is being converted to Standard File Format for search.

3.3.17 Data Base (17): Sociological Abstracts Although not funded by the NSF grant, the UGA Information Center is converting Sociological Abstracts to machine-readable form in a cooperative project with the publisher of this abstracting and indexing service.

4. HARDWARE CONFIGURATION

(The following hardware configuration description contains specifications for the UGA Text Search System which operates within the UGA installation configuration.

The CDC 6400-coupled system is listed since the UT 200 remote printer operates from this computer; search output is automatically routed to this computer.

The complete installation configurations are available on request from UGA.)

4.1 Main Frame:
IBM 360 Model 65
(coupled to CDC 6400)

4.2 Core Size:
Minimum 86K high speed core and 100K Large Core Storage (LCS) bulk core (Georgia is running

with 150-350K high speed and 144K LCS bulk core)

4.3 Mass Storage Devices:
2 9-tr tape drives
1 7-tr tape drive (optional)
2 spindles 2314-equivalent Direct Access Storage Device (varies with volume)
1 drum (optional)
bulk core as described above

4.4 Input Devices
IBM 2260, Uniscope 100, Datapoint 3300 CRT terminals.

4.5 Output Devices:
Same as input devices; IBM 2780, UT 200, and U 9200 remote printers; on-line printers attached to both the 360/65 and the 6400; and off-line printers (IBM 1401s).

5. SOFTWARE CONFIGURATION

5.1 Operating System:
IBM OS, MVT with HASP

5.2 Operational Environment:
All large production jobs, including all searches and conversions, are scheduled on the third shift on the IBM 360/65. Short jobs (less than 5 minutes CPU time) are run in the regular batch job stream during the first two shifts.

Current awareness searches on single issues of the various data bases are converted the same day they are received; the original tape serves as the back-up. The search for that issue is run the following night. Upon completion of the current awareness search, the issue is concatenated with the previous issues for that volume, for the retrospective data base.

Retrospective searches are scheduled over a two-week period. That is, the first volumes of several data bases are scheduled for Monday night of the first week, the next volumes, Tuesday, etc., in order to distribute the search load evenly over the two week period. Jobs which must be rerun due to abends or other problems are scheduled for week-ends unless they can be worked into the nightly schedule.

In FY '71, 557 hours of 360/65 CPU time were

expended on Information Center production work plus 80 hours on research and development work. In addition, 90 hours per month were logged on 1401 computers which are used routinely to print search results.

5.3 Information System:

5.3.1 The Text Search System developed and installed at the University of Georgia Computer Center was designed as a generalized free text retrieval system. The heart of the system is the UGA Text Search retrieval module, developed in 1970. It operates in a batch-mode on data in SFF, the Standard File Format, first developed at the Chemical Abstracts Service (CAS). The system is optimized for the 360, is directory-oriented, table-driven, and operates on variable length records. It uses many CAS macros. Tapes are received from CAS in SDF (Standard Distribution Format) and are converted to SFF on input. Other data bases are also converted to SFF. At the front end of Text Search is an on-line profile creation and maintenance module, OPIUM. This prepares user profiles for SDI current awareness searches and retrospective searches. The total system is a combination of on-line profile management and batch retrieval.

Capabilities and features of the system include:

- On-line profile input and update system
- Character-by-character match technique
- Left and right term truncation
- Threshold weighting
- All data elements are searchable
- Optional card, paper, or tape output
- Uses sequential tape data base files
- Generalized file format
- Table-driven programs

5.3.2 Source Language:
The major programs are written in 360 BAL with utility and conversion programs written in PL/1.

5.3.3 Mode:
Batch retrieval.
On-line profile negotiation and maintenance.

5.3.4 Generalized Packages Used:
The Text-Search System is itself a generalized retrieval system.

5.3.5 Availability:
The program, including data conversion modules, is available for $4000. The documentation price is $100.

6. COMPUTER PROCESSING FUNCTIONS

6.1 Data Definition:
All bibliographic text data bases received by the Center are converted to SFF by conversion programs written for this purpose. Most of the conversion programs do extensive editing and reformatting of the data in addition to changing the file format to SFF. These are individual programs for each data base, written in PL/1.

There are two types of data elements - Left-anchored elements are those types of data for which the format and context are precisely defined. The free text elements are free vocabulary fields in which the context and format are uncontrolled, i.e., abstracts, digests, free-vocabulary index terms and article titles.

6.2 Data Maintenance:
As new tapes are received from the data base suppliers, they are converted by program to SFF. The original issue is retained until after verification of the conversion results and completion of the current awareness search. Both the original tape from the supplier and the converted SFF tape are then concatenated with previous issues to create retrospective volumes for both the original and converted files.

6.3 Data Retrieval:
The Search Program is the heart of the system, and it is designed to operate on Standard File Format data files. Inputs to the program are the profile file and the data base file. The Search Program is table-driven with respect to the data elements, so it references a direct access file of the data elements and their acronyms, identification

numbers, and storage modes. This table, stored as a macro on a private library is recreated or updated as necessary. Output of the Search Program is a tape file of the profiles and the citations which satisfied the search logic, a statistics report on computer time and core utilization for the run, and diagnostics for any errors detected in the profiles.

6.4 Data Output:
Version 2 of the Text Search System, which has been operational since the last quarter of 1971, provides generalized facilities for user-specified output format, content, and media. Normal output in the UGA center is 8 1/2 x 11" narrow paper or 4 x 6" card stock or 11 x 14" wide paper. Remote sites specify their own format. Content can be any combination and arrangement of any data elements contained in the data base. Media include printed or magnetic tape or Direct Access Storage Devices (the latter for remote sites).

6.5 Special:
Profile Management, which is the input portion of the system, uses an on-line, interactive program operating from cathode ray tube (CRT) terminals.

Weighting of search terms (cumulative, non-cumulative and threshold) is a system feature.

At least one backup tape is maintained on all data base tapes.

7. USER INTERFACE

7.1 Current Awareness Service Available?: Yes

7.1.1 Profile Definition:
Information Center staff perform the construction and coding of search profiles after interviewing the customer. The search profiles are entered and updated via an on-line interactive system using CRT terminals. Syntax editing of the profile is done immediately by the computer, with diagnostics returned to the operator of the terminal. The profiles include identification number, threshold weights, Boolean expressions and search terms. There is a maximum of 240 terms per profile. Profile updates are performed as requested.

7.1.2 Output Form:
Standard current awareness and retrospective search output consists of 8 1/2 x 11 paper or cards. The reference contains the title, primary bibliographic citation and any index terms or codes included on the tape. Special output is available on request.

7.1.3 Frequency:
See Table in Section 7.1.4 below.

7.1.4 Charges:
Within the University of Georgia System, Information Center services are considered an overhead item, and are therefore free to university personnel. Searches for others are priced on a cost-recovery basis.

The prices are established on a per-profile per-search run basis. For current awareness searches the price is per-profile per-issue of the data base. For retrospective searches the prices are set on a per-profile per-volume searched basis. Each data base is searched separately, so that one profile searched against two data bases would be invoiced as the sum of the prices for the two files. Partially completed volumes within the current year are billed at the current awareness price (i.e., the issue price times the number of issues in the partial volume).

Commercial and academic users are billed by itemized statement monthly. Searches may be entered or discontinued at any time without penalty.

Subscription discounts are available for organizations who anticipate a large volume of use. In order to obtain the discount the organization must guarantee the minimum subscription. Unused balances are not refundable. Discounts are 5% for subscriptions in the range $2000-$4999 per year, 10% for $5000-$9999, and 15% for $10,000 and above.

CURRENT AWARENESS SEARCHES

(July 1971 - June 1972)

Data Base	Frequency of Updates	Price per Profile to Commercial Organizations
Biological Abstracts (BA)	semimonthly	$ 7
BioResearch Index (BIORI)	monthly	7
CA-Condensates (CA)*	weekly	7
CAIN	monthly	10
Chem.-Biol. Act. (CBAC)	biweekly	5
Compendex	monthly	10
Current Index to Journals in Ed. (CIJE)	quarterly	5
GEO-REF	monthly	7
Nuclear Sci. Absts.	semimonthly	5
Research in Ed. (RIE)	quarterly	5
SPIN	monthly	5
USGRDR	semimonthly	5

*Odd-numbered or even-numbered issues only may be specified.

7.2 Batch Retrospective Search Available?: Yes

7.2.1 Query Formation: The language is the same as for profile definition. Required elements are profile number, threshold weight, Boolean expressions, and at least one search term.

7.2.2 Response Time:
Two or three week turn around schedule.

7.2.3 Charges:
See Sections 2.4 and 7.1.4

7.3 Interactive Retrospective Search Available?: No

7.4 Document Delivery Service Available?: No

8. TRANSFERABILITY CHARACTERISTICS AND EXPERIENCE

8.1 Software Transferability:
The University of Georgia Information Center has experience both as the receiver and donor of software. The Standard File Format (SFF) developed by Chemical Abstracts Service for use in their own internal processing system was the file

structure chosen for the UGA system. Several subroutine packages developed at CAS are incorporated into Text Search. The UGA Information System and its documentation are available for purchase. Information on the use of the system at other installations is not available.

8.2 Data Base Transferability:
All the data bases used in the Center have been purchased (transferred) from suppliers and converted to a single uniform file structure, SFF, which was developed at CAS. These conversion programs are available as part of the software system.

8.3 Methodology Transferability:
The University of Georgia has conducted 90 presentations and on-line demonstrations of its system in FY '71, as well as 32 University of Georgia seminars. Their methodology of converting all data to one common internal format is well known and in use in other installations. They have considerable experience and expertise in choosing data bases, converting data bases and managing a large service-oriented installation.

CATEGORICAL DESCRIPTION OF THE LEADERMART SYSTEM

LEHIGH UNIVERSITY

Prepared for: Office of Science Information Services
National Science Foundation

Prepared by: Systems Development Division
Institute for Computer Sciences and Technology
National Bureau of Standards

January 1973

CONTENTS

CATEGORICAL DESCRIPTION OF THE LEADERMART SYSTEM

LEHIGH UNIVERSITY

1. GENERAL DESCRIPTION

1.1 Project and Organization Name:

LEADERMART
Mart Library
Lehigh University
Bethlehem, Pennsylvania

1.2 Objectives and Operational Philosophy:

LEADERMART is a bibliographic information system operated jointly by the Lehigh Center for Information Science and the Mart Science and Engineering Library. The main objective of the effort is to provide free bibliographic information services to the university community. LEADERMART serves both on-campus and off-campus users, providing access to six data bases, mostly science- and engineering-oriented. The major technical feature is the availability of on-line, conversational access to these data bases.

1.3 Historical Background:

LEADERMART is the product of over 60 man-years of effort which began with Hillman's initial theoretical contributions in 1962. Once the theoretical framework for an automated interactive information retrieval system was established, work began on the construction of a prototype based on the theory. The prototype went through several implementations, and in 1969 the development of the current version of LEADERMART began. It became fully operational in 1971.

1.4 Present Operational Status:

LEADERMART is currently operational and accessible from approximately 30 computer terminals, on and off the Lehigh campus.

1.5 Future Plans:

There are two directions for future work. First, Lehigh expects to continue the creation of specialized data bases. Second, they expect to explore ways of developing knowledge transfer systems for research, instruction and decision-making. The latter effort will feature innovations in information repackaging and the interfacing of data banks with literature bases.

2. ADMINISTRATIVE DETAILS

2.1 Principal Investigator:
Prof. Donald J. Hillman
Director, Center for Information Science
Lehigh University
Bethlehem, Pennsylvania 18015
Phone: (215) 691-7000, x 631

2.2 User Community:
LEADERMART currently serves faculty and students in 12 interdisciplinary research centers at Lehigh, seven local colleges, the Environmental Protection Agency, the National Bureau of Standards, the Environmental Data Service of NOAA, the Naval Ships Research and Development Center, and the Naval Air Development Center. Negotiations are underway with a number of outside organizations, especially chemical and pharmaceutical companies and member firms of the New York Stock Exchange.

2.3 Documentation:

(1) Donald J. Hillman, "Negotiation of Inquiries in an On-line Retrieval System:, Information Storage and Retrieval, Vol. 4, pp. 219-238, Pergamon Press, 1968.

(2) Donald J. Hillman and Andrew J. Kasarda, "LEADERMART System and Service," paper presented at ACM National Conference, August, 1972, available from Lehigh.

(3) "LEADER Information Manual", document dated November 14, 1972, available from Lehigh.

(4) "The LEADERMART Project", collection of selected progress reports to NSF, available from Lehigh.

(5) "Interactive Chemical Information Retrieval via LEADERMART", paper presented at ACS National Conference, available from Lehigh.

2.4 Services and Costs:
Service is free to Lehigh users. Use of the on-line facility costs $55.00 per connect hour to outside users ($45.00 per connect hour to government users).

3. DATA BASES

3.1 Data Base Selection:

3.1.1 Requirements and Acquisition Strategy:
Currently searchable data bases were selected (or created) on the basis of the needs of Lehigh's users.

3.1.2 Discipline Coverage:
Science and engineering subject matters are currently covered, together with cataloging information via MARC II.

3.2 Data Preparation, Conversion and Entry:
Tapes from commercial suppliers are converted to LEADERMART's format. In some cases tapes are edited and in some cases (where the vocabulary supplied is not adequate for use by LEADERMART) tapes are run through a text analyzer for indexing.

3.3 Data Base Contents:
About 300,000 titles and abstracts are accessible on-line. At least 18 recent months of recurring data bases (e.g., CA-Condensates) are held on-line. The thesaurus of index terms numbers 1.4M terms, and is growing at 8000 terms per month.

3.3.1 Data Base (1): ASCE Journal Abstracts (Civil Engineering).
These are bibliographic citations extracted from American Society of Civil Engineering Journals.

3.3.2 Data Base (2): CA-Condensates
CA-Condensates is the computer searchable complement to the printed publication, Chemical Abstracts (CA), which covers the full range of chemistry, referencing 250,000 articles per year.

3.3.3 Data Base (3): Engineering Index Abstracts: Compendex.
Compendex is the computer-readable version of the printed publication which contains references spanning all engineering disciplines.

3.3.4 Data Base (4): Tall Structures Abstracts
This is a specialized data base, created locally. It currently consists of 216,000 references.

3.3.5 Data Base (5): MARC II
MARC II, a data base containing bibliographic citations prepared by the Library of Congress, is accessible via LEADERMART, which is an on-line, interactive cataloging system.

3.3.6 Data Base (6): School of Information Science Documents. About 300 documents in full text form constitute this data base, which is used only for experimental purposes, especially in testing the text analyzer.

4. HARDWARE CONFIGURATION

4.1 Main Frame:
CDC 6400

4.2 Core Size:
65K, 60 bit words.

4.3 Mass Storage Devices:
One Billion-character disk.

4.4 Input and Output Devices:
Teletype terminals as well as other terminals of higher speeds such as 300 and 1200 baud capacity can be served. By January 1973, 4800 baud capacity will be available.

5. SOFTWARE CONFIGURATION

5.1 Operating System:
The system operates under the SCOPE 3.3 operating system and the locally developed INTERCOM 3.0 time sharing system.

5.2 Operational Environment:
LEADERMART is run on the CDC 6400, which is the University's only computer, processing a typical university workload.

5.3 Information System:

5.3.1 Name and Brief Description:
LEADERMART consists of four system components controlled by LEADER, the monitor. The four components are: (1) information entry, editing, and text processing, (2) connectivity, file generation, and

maintenance, (3) retrieval, and (4) evaluation and feedback.

5.3.2 Source Language:
The programs are primarily in FORTRAN.

5.3.3 Mode:
The predominant use is on-line. For each LEADERMART user, a 25K core partition is required since the retrieval package is non-reentrant. The Lehigh computer accepts about 12-15 concurrent users, hence access to LEADERMART service is quite limited during peak computer usage periods.

5.3.4 Generalized Package Used:
None

5.3.5 Availability:
Available, but no cost breakdown has been worked out yet. Certain disk input-output routines are considered proprietary.

6. COMPUTER PROCESSING FUNCTIONS

6.1 Data Definition:
There is the possibility of creating specialized data bases. The "Tall Buildings" data base is an example. Text entry and file creation is performed as a batch operation.

6.2 Data Maintenance:
There are, in addition to the editing and indexing previously mentioned, the usual updating activities.

6.3 Data Retrieval:
A search request in natural English prose is directed to a logico-syntactic analyzer which determines what the request is about. Connectivity files then give a statement of the interconnections among the systems terms and a statement of the association of terms with documents in the collection. Using a ranking algorithm based on the analysis of the request, the characteristics which are related to the request are then presented to the user. After optional re-negotiations, which may widen or narrow the search, a document list associated with those characteristics is presented.

6.4 Data Output:
Intermediate data is presented via a cathode ray tube terminal. Final document list is presented either via CRT terminal or high speed printer.

6.5 Special:
There are extensive browsing and tutorial features. System can completely monitor and record each user's interactive dialogue.

7. USER INTERFACE

7.1 Current Awareness Services Available?: Yes

7.1.1 Profile Definition:
Profile definition assistance is available from information specialists in the MART Library.

7.1.2 Output Form:
Hard copy is disseminated periodically. Optionally, the system will update a user's files automatically as information captured by the profile comes in.

7.1.3 Frequency:
Monthly

7.1.4 Charges:
$15.00 per month per profile, regardless of data base.

7.2 Batch Retrospective Search Available:
Yes, but the furthest back any data base goes is 1970 (for CA Condensates).

7.2.1 Query Formation:
User provides a descriptive paragraph (or alternatively, fills out a form) and lists special terminology.

7.2.2 Response Time:
Two weeks, maximum.

7.2.3 Charges:
COMPENDEX: $100.00 per profile against two years,
CONDENSATES: $200.00 per profile against two years,

plus $1.00 per document for first 100 documents,
$0.50 per document for next 400 documents,
$0.25 per document thereafter.

7.3 Interactive Retrospective Search Available?: Yes, but covering at most 24 recent months of recurring data bases.

7.3.1 Query Language:
Natural English prose.

7.3.2 Training and Assistance:
Both are available. Furthermore the system is highly tutorial.

7.3.3 Response Time:
Almost immediate. This currently depends on how many other users are using LEADERMART.

7.3.4 Charges:
$55.00 per connect hour ($45.00 for government and universities.)

7.4 Document Delivery Service Available?: Yes, on special request.

7.4.1 Form of Delivery:
Xerox copies are delivered.

7.4.2 Charges:
Charges cover reproduction and postage costs.

8. TRANSFERABILITY CHARACTERISTICS AND EXPERIENCES

8.1 Software Transferability:
The software (which is almost completely in FORTRAN) is flexible and hence relatively transferable. Various program modules have been run on a 360/65 and only I/O statements had to be changed.

8.2 Data Base Transferability:
There would be no problem in transferring the data bases to any machine. The reformatting of commercial data bases is a minor matter.

8.3 Methodology Transferability:
Since the methodology is so heavily natural language oriented and machine-independent, there is no problem in methodology transfer. It would be far easier to convert the software than to start programming from scratch with a view towards implementing the LEADER methodology.

CATEGORICAL DESCRIPTION OF OHIO STATE UNIVERSITY MECHANIZED INFORMATION SYSTEM

Prepared for: Office of Science Information Services
National Science Foundation

Prepared by: Systems Development Division
Institute for Computer Sciences and Technology
National Bureau of Standards

January 1973

CONTENTS

CATEGORICAL DESCRIPTION OF OHIO STATE UNIVERSITY

MECHANIZED INFORMATION SYSTEM

1. GENERAL DESCRIPTION

1.1 Project and Organization Name:

The Mechanized Information Center (MIC), Ohio State University

1.2 Objectives and Operational Philosophy:

The goal of MIC is to provide information services that make use of machine-readable bibliographic and other data bases. These services are to be provided mainly for students and faculty of the Ohio State University. MIC has set three objectives which will lead to the achievement of its goal:

(1) To acquire bibliographic data bases

(2) To attempt to maximize the service potential of the data bases

(3) To encourage user-oriented research into the problems of information centers.

MIC is administratively a part of the Public Services Division of the University Libraries. The internal organization of MIC consists of three major functional areas:

(1) The information area, responsible for developing and maintaining patron profiles,

(2) The operation area, concerned with the day-to-day operation of MIC,

(3) The programming area, responsible for the development and maintenance of the software.

The unique aspect of MIC is its utilization of an integrated discipline-crossing data base. Since October 1971 current awareness service from the multi-disciplinary data bank have been provided to 1185 patrons on a bi-weekly basis. Retrospective searches from the multi-disciplinary data bank will be implemented during the fall of 1972.

1.3 Historical Background:
On February 1, 1971, MIC received a grant (Gr-27458) from the Office of Science Information Services of the National Science Foundation to help support MIC activities during the first year of a four-year developmental period. During the first year, MIC has recruited a staff of systems designers, librarians and information scientists and set up a three-part organization. MIC, although jointly sponsored by the Department of Computer and Information Science and the University Libraries, is administratively a part of the Public Services Division of the University Libraries. The Director of MIC reports to the Assistant Director of Libraries, Public Services, and holds a joint appointment in the Department of Computer and Information Science. During the first year MIC implemented the current awareness service from the multi-disciplinary data bank, which contains bibliographic information about all journal articles appearing in 3,200 journals, government reports and books. The current awareness service was officially started on October 22, 1971. The service began with 36 profiles and grew to 285 profiles as of January 31, 1972. Retrospective searches from the multi-disciplinary data banks will be implemented in October 1972. There were 5,367 searches made in the period ending January 31, 1972.

1.4 Present Operational Status:
The current awareness service is operational since October 1971 and currently serves 1185 patrons (August 1972).

1.5 Future Plans:
Retrospective services from the multi-disciplinary data bank will be offered during the fall of 1972. Future plans consist of evaluation and refinement of Multi-disciplinary Services and testing and implementation of a discipline-based current awareness service to provide in-depth coverage in each of four areas: biological sciences, physical sciences, engineering, and social sciences. Areas of research consist of an automatic profiling system, a systems evaluation system, a systems evaluation of MIC in terms of user-satisfaction and relevance of service, a testing and evaluation of TEXT-PAC, a software package available from IBM, and the development of a management information system based on MIC's operation. Another research

project is the marketing program to market the MIC service.

2. ADMINISTRATIVE DETAILS

2.1 Principal Investigator:

Dr. Gerald Lazorick, Associate Professor
Department of Computer & Information Science
The Ohio State University
Columbus, Ohio 43212
(614) 422-3480

2.2 User Community:

Faculty and graduate students of Ohio State University and other organizations such as EPA, Battelle, General Motors, etc.

2.3 Documentation:

Annual Report of the Mechanized Information Center, February 1971 through January 1972.

2.4 Services and Costs:

MIC services are free to Ohio University faculty and graduate students. The charge for current awareness services on the multi-disciplinary data base is $300 per profile per year with a minimum of 5 profiles for non-university users. The average charge for the retrospective search service on multi-disciplinary data base will probably be $25 for non-university users.

3. DATA BASES

3.1 Data Base Selection

3.1.1 Requirements and Acquisition Strategy:

The goal of MIC is to provide information services that make use of machine-readable bibliographic and other data bases. These services are to be provided mainly for the students and faculty of the Ohio State University. Data base overlap studies were conducted to determine what data base might, or might not, be a valuable addition as a disciplinary data base. This process also serves as a justification for the present multi-disciplinary data base through step-by-step comparison with smaller, more easily analyzed data bases. Detail of overlap studies is documented on page 25 of Annual Report 1971-1972.

3.1.2 Discipline Coverage:
The multi-disciplinary data bank at present is a composite of four individual data bases merged into one. The coverage is heaviest in the physical and life sciences, and in technology. Types of coverage in the data base include:

(1) Journal literature

(2) Government reports

(3) Monographs

As of January 31, 1972 it contained more than 90,000 citations.

MIC also will be having discipline-based data bases in each of four areas: biological sciences, physical sciences, engineering, and social sciences. Possible data banks for these areas include: Chemical-Biological Activities (CBAC), Computerized Engineering Index (COMPENDEX), Geological File (GEO-REF), and many other specialized data bases in machine-readable form. MIC is already offering Chem Title (CT) service to faculty members who have a specific interest in chemistry-oriented journals.

3.2 Data Preparation, Conversion and Entry:
MIC uses PANDEX vocabulary as the basic word list. Tapes are received from the supplier and conversion programs are run to reformat into an internal format. MIC decided to use PANDEX as the basis for internal format. The conversion programs read data tapes and create two output files:

(1) A word record tape containing significant words with article numbers.

(2) A random access article file containing title, author, and journal information ordered by article number.

The word record file is then sorted alphabetically by word, and sequentially by article number. This sorted record tape is then input to the inverted file generation program, which creates a random access inverted file by words and article number. (See schematic diagram on p. 62 of First Annual Report).

3.3 Data Base Contents:
The multi-disciplinary data bank consists at present, of four individual data bases merged into one with duplicates eliminated. These, and one discipline-oriented data base, are described below.

3.3.1 Data Base (1): PANDEX journal
The PANDEX tape contains bibliographic information on the articles and technical notes appearing in more than 2400 journals. The bibliographic information includes the title of the articles, name(s) of the author(s), name of the journal, Coden abbreviation, issue number, page number, and subject heading taken from a Pandex thesaurus. In one year, more than 550,000 citations have been collected in the MIC multi-disciplinary data bank. Tapes are issued weekly.

3.3.2 Data Base (2): ISI journal
The ISI source tapes are issued weekly by the Institute for Scientific Information which is a commercial organization in Philadelphia.

The ISI Source tapes are run against a conversion program to delete those journals that are already on the Pandex tapes and to change the format of the items that remain to correspond to the format of the Pandex tapes.

In addition, the original ISI tapes contain such peripheral items as reviews, editorials, letters. The MIC conversion program deletes these.

The information for each article or note includes: title of the article, author(s), an abbreviation of the journal name, volume, issue number, and page number.

The journal abbreviations are special 11-character sets of letters devised by ISI; they are not Coden. In addition, there are no subject headings.

The Source index tapes include items from foreign journals. These titles are translated into English and preceded by a two-character code to indicate the language of the article, if not English. Since the service began, 12 tapes have been converted and searched. After conversion into the PANDEX format, the tapes yielded more than 25,000 citations.

3.3.3 Data Base (3): NTIS

NTIS is the National Technical Information Service of the U. S. Department of Commerce. It consists of unclassified government reports resulting from government-sponsored research. The reports are compiled in a publication now called Government Reports Announcements (GRA). GRA is available in machine readable form from the Department of Commerce. However, MIC receives the tapes through CCM Information Corporation, which reformats the original tapes into standard Pandex format. In one year, 40,000 to 50,000 reports are indexed by NTIS and appear as part of GRA. The information is in 22 fields, mainly science, engineering and mathematics, but also includes reports on behavioral and social sciences. More than 20,000 citations are already in the MIC Data Bank.

The tapes contain standard bibliographic information such as author and title, as well as abstracts, descriptor terms, and information on how to order copies. Tapes are received and searched twice a month by MIC. So far, six tapes, which contain 20,000 citations have been searched.

3.3.4 Data Base (4): MARC

Later this year, MIC expects to add MARC tapes to its data bank. These tapes contain bibliographic material on more than 100,000 books cataloged by the Library of Congress. The fields covered are the hard sciences, social sciences and technology.

MIC expects to obtain these tapes through the Ohio College Library Center (OCLC), which is on the OSU campus. Preliminary discussions are now underway between MIC and OCLC.

3.3.5 Data Base (5): Chemical Titles
This is not merged into the multi-disciplinary data bank. It is a discipline-oriented data base containing papers from approximately 730 journals in the field of chemistry and chemical engineering. Chemical Titles (CT) presents the titles of papers published in journals before an abstract of the article appears in Chemical Abstracts. The information on each CT paper includes title, author, and complete citation of paper, and is in machine-readable form on magnetic tape. The tapes are issued bi-weekly and furnish approximately 130,000 citations in one year.

4. HARDWARE CONFIGURATION

4.1 Main Frame: IBM 370/145

4.2 Core Size: The System requires a 512K machine to run. The nucleus software called MIC SEARCH occupies 160 to 180K of core.

4.3 Mass Storage Devices: Minimum of six 3330 disk drives (3 for OS and 2 for MIC SEARCH), and 3 tape drives.

4.4 Input Devices: Card reader, tape units

4.5 Output Devices: Printer

5. SOFTWARE CONFIGURATION

5.1 Operating System:
IBM OS with MVT and HASP.

5.2 Operational Environment:
The execution of MIC SEARCH requires the total machine because of the disc pack requirements.

5.3 Information System

5.3.1 Name and Brief Description:
The current MIC current awareness system was built around programs obtained from the Technical Information Dissemination Bureau (TIDB), SUNY at Buffalo. The system consists of ten application programs written in-house and three programs supplied by IBM. Inputs to the system are profiles that reflect the user's areas of interests and the MIC data bank. The primary output of the system is a set of notifications containing descriptions of those articles, reports, and papers likely to be of interest to each user of the system.

5.3.2 Source Language:
The Search program is written in PL/1 with some assembler language subroutines. The print program is written in COBOL.

5.3.3 Mode of Operation:
Batch for search, maintenance and profile definition.

5.3.4 Generalized Packages Used:
MIC SEARCH is operable as a package on any IBM 360 or 370 (OS) computer. Input and output formats are fixed but can be modified.

5.3.5 Availability:
The MIC SEARCH software is available, but MIC does not have a price or policy for distribution. They have furnished the CCM Information Corporation with the software in exchange for the ERIC tapes (RIE and CIJE).

6. COMPUTER PROCESSING FUNCTIONS

6.1 Data Definition:
There is no data definition capability. Conversion programs were written to reformat the input data record into an internal format. MIC decided to use the PANDEX format as the internal one for all data banks. The PANDEX format is described on page 111 Appendix B of the same Annual Report.

6.2 Data Maintenance:
Data Maintenance consists only of adding new

records. No modification can be made to the existing files except those changes that are reflected in the suppliers tapes. The mode of operation is batch and there are weekly runs for PANDEX and ISI, and MARC, and twice a month for NTIS.

6.3 Data Retrieval:
Data searching is done on the inverted files. The inverted files are organized for direct access with word list followed by list of article numbers which translates into relative address for the master record. The search terms or profiles consist of a series of words or groups of words with a term weight that reflects the probability that a user would be interested in a citation that has that word or word group in its title. Each title that contains words that match with one or more profile terms, yields a total significance value. When the total significance value of an article is equal to, or greater than, a particular threshold value for the user profile, a notification describing the citation is generated. The search logic is described on page 38 of the Annual Report.

6.4 Data Output:
There are no facilities for user-specified output form. The output of the current awareness search is a set of notification cards. The cards contain citation information on the left and a stub on the right containing an indication of the location of the journal in a campus library. The stub also becomes the order form for copy service. The outputs are mailed out every other Friday.

6.5 Special:
There are no special functions such as input editing, input validation, data security feature, etc. There is no restart and recovery procedure, but 75 tapes are maintained of the backup system programs decks and profiles. There is a profile maintenance program.

7. USER INTERFACE

7.1 Current Awareness Service Available?: Yes

7.1.1 Profile Definition:
The profiles are defined by the user with the assistance of an information specialist of MIC. They are then coded, keypunched

and kept on the profile tape. The initial output of a current awareness search is screened for monitoring purposes. After 2 or 3 months, the user is recontacted for an update of the profile.

7.1.2 Output Form:
The output consists of 3 x 5 notification cards which are two-part units, the left side containing citation information and the right side containing reference location information for the user.

7.1.3 Frequency:
Current Awareness Searches are run every two weeks.

7.1.4 Charges:
Services are free to the university user, but there is a charge to non-university user of $300 per profile per year.

7.2 Batch Retrospective Search Available?: Not yet, but the service will be available as soon as software is completed.

7.3 Interactive Retrospective Search Available?: No.

7.4 Document Delivery Service Available?: Yes.

7.4.1 Form of Delivery:
MIC is offering optionally a first page copy service which would usually contain an abstract of an article.

7.4.2 Charges:
At present the charge is 10¢ for each first page.

8. TRANSFERABILITY CHARACTERISTICS AND EXPERIENCES

8.1 Software Transferability:
The MIC SEARCH software was transferred from State University of New York at Buffalo. Modifications and extensions were made and the software was made operational in a relatively short time (3 months with 2 programmers and 1 programming manager). The software also includes 3 utility-programs supplied by IBM which presumably are generalized. MIC staff says that MIC software has been distributed to the CCM Information Corporation, but no

further elaboration was given.

8.2 Data Base Transferability:
The disciplined data base (Chemical Titles) is kept in the supplier's format and is therefore transferable with CT format software.

The multi-disciplinary data base under OS consists of a series of logical records written on a physically sequential medium. The logical records are unblocked and varying length but are not in IBM Variable format. Therefore, they must be read in IBM undefined format.

8.3 Methodology Transferability:
MIC has conducted a data base overlap study. A survey of major mechanized data bases and the discipline as well as each journal covered were analyzed. The study is documented in Section 2.3, page 25 of the Annual Report 1971-1972. The survey of data bases appears in Appendix 1, page 203 of the same Annual Report.

The search logic and inverted file structuring of TIDB software is straight forward and the methodology is already well known.

MIC has done some work in data inputting costs. The methodology is described in Appendix F, page 136 of the Annual Report. Some marketing effectiveness questionnaires are being used.

CATEGORICAL DESCRIPTION OF CAMPUS-BASED INFORMATION SYSTEM

UNIVERSITY OF PITTSBURGH

Prepared for: Office of Science Information Services
National Science Foundation

Prepared by: Systems Development Division
Institute for Computer Sciences and Technology
National Bureau of Standards

January 1973

CONTENTS

CATEGORICAL DESCRIPTION OF CAMPUS-BASED INFORMATION SYSTEM

UNIVERSITY OF PITTSBURGH

1. GENERAL DESCRIPTION

1.1 Project and Organization Name:

Campus-Based Information System (CBIS)
University of Pittsburgh
Pittsburgh, Pennsylvania

1.2 Objectives and Philosophy:

Discipline or problem-oriented information services representative of advanced computer technology are being introduced and related to the University Library system. Services are obtained commercially from Government, professional societies or other universities. The intent is to develop a "one-stop" service providing an individual user with one entry point to all data bases and specialized services relevant to his interest. Batch-oriented and interactive computer searching would be available, but users are expected to rely heavily on comprehensive assistance from information specialists. Research efforts emphasize the analysis of information needs, usage patterns, and distribution patterns among user populations, presenting a further role for information specialists.

1.3 Historical Background:

Pittsburgh has actively pursued the development of "knowledge availability systems" since the early 1960's. Noteworthy antecedents of the CBIS project are the NSF-funded Pittsburgh Chemical Information Center (1967-70) and the NASA/industry funded Knowledge Availability Systems Center (KASC) which has served operationally as a NASA Regional Dissemination Center since 1964. KASC provides "batch" searches for industry and other users of NASA research bibliographic files.

1.4 Present Operational Status:

As of August 1972, a new search software package (PIRATES - Pittsburgh Information Retrieval and Text Search) has become operational on Digital Equipment Corporation PDP-10 computers at the Pittsburgh Computer Center. This is now used for batch service through KASC, or can be operated at

a remote terminal for interactive searching, which is presently limited to small files, e.g., one month's issue of the American Society of Metals METADEX file.

An Information Utilization Laboratory (IUL) has been operated in the Engineering Library since October 1971, including an interactive computer terminal. In September 1972, an IUL in Social Science will be initiated. (Two additional IUL's in Humanities and Medicine are planned for 1972, not involving NSF funds).

A computer terminal to the New York Times data bank is expected to become operational in September 1972, to function in conjunction with the Social Science and Humanities IUL's, or perhaps to become the central facility of a new IUL.

1.5 Future Plans:
The period of NSF support is considered the conceptual phase of CBIS development, addressing feasibility studies and initial system design. During 1972-3, the interface between users and information specialists will be extended by establishment of more discipline-oriented Information Utilization Laboratories, the "one-stop" service centers. Computer-based capabilities at Pittsburgh are being expanded with additional data bases, improved search software, and special facilities such as computer-output microfilm. Resource-sharing arrangements of a broad variety are being pursued, within the Pittsburgh vicinity and nationally. A Central Analysis Unit will be initiated in September 1972 to undertake collective studies of information utilization patterns as documented in the several operating IUL's.

2. ADMINISTRATIVE DETAILS

2.1 Principal Investigator:
Professor Allen Kent
Director, Office of Communications Programs
Library and Information Science Building
135 North Bellefield
Pittsburgh, Pennsylvania 15213
Telephone: (412) 621-3500, extension 6352

2.2 User Community:
The CBIS research is primarily focused on

Pittsburgh faculty at the present time, through the new Engineering and Social Science IUL's. However, the concept addresses all segments of the intellectual community, so research and service among students for example, may expand in the near future. The KASC provides current awareness search service to 104 clients on an annual contractual basis, with over 500 individual profiles.

2.3 Documentation:

(1) I Annual Report (in lieu of IV Quarterly Report) February 1, 1971 to January 31, 1972. University of Pittsburgh Campus-Based Information System to National Science Foundation, Grant No. G-27537.

(2) Quarterly Progress Report V, February 1 to April 30, 1972. University of Pittsburgh Campus-Based Information System to National Science Foundation, Grant No. G-27537.

(3) Quarterly Progress Report VI, May 1 to July 30, 1972. University of Pittsburgh Campus-Based Information System to National Science Foundation, Grant No. G-27537.

(4) "A Campus-Based Information System at the University of Pittsburgh", Office of Communications Programs, University of Pittsburgh, June 1, 1970.

(5) "PIRATES and the ASM File", Unpublished draft available from Office of Communications Programs, University of Pittsburgh.

2.4 Services and Costs:
See attached Fee Schedule, KASC Information Services, May 30, 1972, pertinent to industrial clients.

Services are provided without direct charge to Pittsburgh faculty and students.

3. DATA BASES

3.1 Data Base Selection:

3.1.1 Requirements and Acquisition Strategy: Machine-readable data bases are obtained from not-for-profit and profit-oriented commercial sources. Selection decisions

reflect assessment of costs in relation to the potential marketability among campus users and industrial clients of KASC, but also include consideration of maximum effectiveness for users in information seeking. For example, two commercial data bases are being acquired from the Institute for Scientific Information for user comparison with the specialized data bases already available. ERIC files are being strongly considered for acquisition.

3.1.2 Discipline Coverage:
Discipline coverage for CBIS is intended to be comprehensive, through locally held data bases or by remote access to such services as the New York Times data bank, MEDLINE and MEDLARS. KASC provides service on certain data bases, e.g., COMPENDEX, through other universities participating in the NASA dissemination project.

3.2 Data Preparation, Conversion and Entry:
Pittsburgh search software (PIRATES) utilizes one common input format, and all acquired data bases must be converted. Conversion delays have caused temporary interruptions to standard services, such as retrospective search on NASA files.

Two conversion steps are involved. Since most data files are received in an IBM-oriented format (e.g., EBCIDIC code), they are first converted to a standard text tape format according to DEC PDP-10 conventions. The second step converts these tape files into the PIRATES standard search format, which involves a linked-list organization of all words in a text record, suitable for a binary search against a set of user profile terms.

3.3 Data Base Contents:
The KASC fee schedule indicates the available data bases, described more fully below. Regarding CA-Condensates, Pittsburgh has found the five major subject categories particularly suited to its user population. Profile studies have shown that one subject category will provide 75% of the relevant citations in the entire file, while two selected categories can provide 90% of relevant documents.

The Library of Congress MARC file is being used experimentally in conjunction with computer-output

microfilm equipment. The CBIS project foresees this as a means to economically provide specialized card catalogs on microfilm for IUL's and special library centers.

3.3.1 Data Base (1): CA-Condensates (CA-C)
CA-C is the computer searchable complement to the printed publication, Chemical Abstracts (CA), which covers the full range of chemistry, referencing 250,000 articles per year.

CA-Condensates is issued weekly; the content corresponds to an issue of CA. The tape version, CA-Condensates, precedes the corresponding printed issue of CA by several weeks due to the time required to print, bind, and distribute CA printed issues.

The abstracts in CA and CA-Condensates are grouped into five categories: Biochemistry, Organic Chemistry, Macro-molecular Chemistry, Applied Chemistry and Chemical Engineering, and Physical and Analytical Chemistry. The first two groupings are published as an odd numbered issue one week, and the last three groupings are published as an even numbered issue the following week. Searches may be limited to odd or even numbered issues if desired.

Pittsburgh has available seven volumes, beginning July 1968.

3.3.2 Data Base (2): Chemical Titles (CT)
CT, which is issued by Chemical Abstracts Service, contains journal references to approximately 4,500 articles per issue appearing in 650 important U. S. and non-U. S. chemical and chemical engineering journals. Titles that appear in CT represent over 65% of the total abstracts that later appear in Chemical Abstracts. Chemical Titles offers journal references to articles approximately 70 days before their abstracts are published in Chemical Abstracts. In many cases titles appear in

Chemical Titles before the journal containing the article is published. Thus Chemical Titles is valuable as an alerting service.

3.3.3 Data Base (3): Computerized Engineering Index (COMPENDEX). COMPENDEX, issued monthly by Engineering Index, Inc., is the computer-readable version of the printed publication, Engineering Index Monthly, which contains references spanning all engineering disciplines. These references are taken from professional and trade journals, publications of engineering organizations, papers from conferences and symposiums, and books and other documents.

This data base is made available at Pittsburgh through Indiana University (batch service only).

3.3.4 Data Base (4): GRA
Government Reports Announcements is available in machine readable form from the Department of Commerce. It consists of unclassified government reports resulting from government-sponsored research.

3.3.5 Data Base (5): NASA File
The NASA File consists of the STAR and IAA documents. Tapes are received monthly containing approximately 4,500 documents.

3.3.6 Data Base (6): ASM/IM
The American Society of Metals' METADEX file issued every month contains approximately 1,600 entries.

This data base is made available at Pittsburgh through the University of Connecticut, an associate center in the NASA regional dissemination activity.

4. HARDWARE CONFIGURATION

4.1 Main Frame:
Digital Equipment Corporation PDP-10 dual processor (KI-10) configuration.

4.2 Core Size:
256K words per processor

4.3 Mass Storage Devices:
15 RP03 magnetic disk drives, 51M char. each.

4.4 Input-Output Devices:
Data terminals must be ASR-33 teletype compatible.

5. SOFTWARE CONFIGURATION

5.1 Operating System:
Standard time-shared operating system for PDP-10 dual processor offered by DEC.

5.2 Operational Environment:
Multiprogrammed, with foreground time-sharing and background batch processing.

5.3 Information System

5.3.1 Name and Brief Description:
PIRATES is a full text search system developed at Pittsburgh, based upon original design concepts and aimed at effectively utilizing the particular capabilities of the PDP-10, including interactive time-sharing. Users may specify search terms of 24 characters maximum length, embedded blanks being prohibited. Left and/or right truncation may be used to construct a stem. Boolean AND, OR, NOT logic may be used to combine terms in a profile, and a connector operator is provided to concatenate single terms into multiple word phrases. The types of data to be searched for a match (e.g. title, author, journal citation, keywords) may be selected in various combinations.

The profile specification statements have a card-oriented format (e.g. columns 73-80 a are not used, and all statements must begin with a proper character, not a blank, in column 1) whether entered from a console or via a card deck. The software gives minimal guidance and cues to the console user who is asked to answer a few questions to select various options (e.g. DO YOU WISH DOCUMENTS DISPLAYED?) and is given the

several possible acceptable answers, (e.g. ANSWER "YES", "NO", "HEAD", OR "LEVELS".) After examining potential retrievals from on-line searching of a small file, the console user may restart the profile entry sequence to refine his profile definition.

5.3.2 Source Language:
PIRATES is implemented in PDP-10 assembly language, using available macros of the operating system.

5.3.3 Mode:
The basic search strategy is to serially pass all document records in a file against the collected set of profiles, looking for a match with any of them. The system operates using about 10K words of core memory during the search.

5.3.4 Generalized Packages Used:
Exclusing operating system macros and utility routines, the system uses no major generalized software packages. PIRATES is modularized into a basic search package and a "co-routine controller" which governs input-output and the order of functions performed.

5.3.5 Availability:
Although partially developed with University funds as well as NASA and NSF funds, PIRATES is available with minimal charge for reproduction of tapes, decks, and existing documentation.

6. COMPUTER PROCESSING FUNCTIONS

Development and improvement of PIRATES is continuing, and documentation on its design has not been completed. Certain design attributes are reported from conversation with the principal designer, Professor Dale Isner.

6.1 Data Definition:
Searchable files are created by one program, CONVERT. The program accepts input text on-line from a data terminal. Alternatively, a text file may be created on-line using the PDP-10 system

text editor, and then passed as input to CONVERT.

6.2 Data Maintenance:
Two already converted files may be concatenated by a system program, and this is the usual procedure for adding new documents to a data base. Corrections would be made in the original text files.

6.3 Data Retrieval:
The basic retrieval strategy is a serial search of the subject data base.

6.4 Data Output:
Complete document information or selected portions may be printed on-line or may be set up for off-line printing.

6.5 Special Functions:
Data security, recovery procedures, editing, and utility functions are largely determined by PDP-10 operating system capabilities.

7. USER INTERFACE

7.1 Current Awareness Service Available? Yes

7.1.1 Profile Definition: On-line by user or through KASC assistance for off-line input.

7.1.2 Output Form: Printed listing

7.1.3 Frequency: Monthly to weekly, depending upon data base.

7.1.4 Charges: See fee schedule, Exhibit 1

7.2 Batch Retrospective Search Available? Yes

7.2.1 Query Formation. On-line or through KASC assistance.

7.2.2 Response Time: As rapidly as manual handling and machine availability permit.

7.2.3 Charges: See fee schedule.

7.3 Interactive Retrospective Search Available?
Yes, for sufficiently small data bases permitting adequate response time and efficient machine utilization.

7.3.1 Query Language: Profile definition language of PIRATES.

7.3.2 Training and/or Assistance: Through IUL staff and user-oriented documentation, at present.

7.3.3 Response Time: Not observed.

7.3.4 Charges:
No direct charge to University users; not known for external users.

7.4 Document Delivery Service Available? Yes.

7.4.1 Form of Delivery: Abstract of CA-Condensates delivered by mail from KASC. Other documents may be provided by IUL personnel.

7.4.2 Charges: 15¢ each for CA abstracts.

8. TRANSFERABILITY CHARACTERISTICS AND EXPERIENCES

8.1 Software Transferability:
PIRATES would be directly transferable to a single processor PDP-10 with similar peripheral devices. Being completely machine dependent, software is not transferable to other computers.

8.2 Data Base Transferability:
Limited by PDP-10 dependence and unique search software.

8.3 Methodology Transferability:
Undetermined pending design documentation availability.

Exhibit I

May 30, 1972

THE KNOWLEDGE AVAILABILITY SYSTEMS CENTER

UNIVERSITY OF PITTSBURGH, PITTSBURGH, PENNSYLVANIA 15213
PHONE 621-3500

FEE SCHEDULE

KASC INFORMATION SERVICES

The KASC data base now includes over a million items and is growing at a rate of more than 43,000 items each month. The following is a list of the component files, the search services available from each, and the fee schedules.

NASA FILE

- Custom Profiles
  - Current Awareness Service (12 consecutive months)
    - With Engineering Review $180
    - Without Review $ 85
  - Retrospective Search Service (1962 to date)
    - With Engineering Review $185
    - Without Review $ 90
  - Retrospective and Current Awareness
    - With Engineering Review $275
    - Without Review $135
- Standard Interest Profiles
  - Current Awareness Service (12 consecutive months) $ 96

CHEMICAL ABSTRACTS SERVICE FILES

Custom profiles for current awareness service are available from the following:

- CAS CONDENSATES
- CAS CHEMICAL TITLES

The fee for either service is $50 to enter a profile,

plus $5 per search per profile. The entry fee includes a review of the results of the first three searches and client liaison in order to optimize the search strategy. Client-requested strategy changes after the first three searches are assessed at $10 additional.

Retrospective searching of CAS CONDENSATES for custom profiles is available at $50 to enter a profile, plus $5 per volume per section. Currently there are seven volumes available beginning with July, 1968, Volumes 69 through 75, and each volume consists of five sections corresponding to the five major subject groupings of the Chemical Abstracts journal. The $50 fee is assessed only once per profile regardless of the type of service provided: current awareness, retrospective, or a combination of both.

ASM/IM METALS ABSTRACTS

- Custom Profiles
  - Current Awareness Service (12 consecutive months) $215
  - Retrospective Service (1966 to date) $185
  - Retrospective and Current Awareness $365

Fees include all start-up costs and a review of the results of the first three searches and client liaison in order to optimize the search strategy. Client-requested strategy changes after the first three searches are made at no charge to the client.

GOVERNMENT REPORTS ANNOUNCEMENTS*

- Custom Profiles
  - Current Awareness Service (12 consecutive months)
    - With Engineering Review $180
    - Without Review $ 85

*Formerly United States Government Research and Development Reports (USGRDR). Retrospective service on this file for the period August 1964 to September 1970 covers only the unclassified portion of the Defense Documentation Center File. Beginning in September 1970 the unclassified DDC file was incorporated into USGRDR and continues to be included with Government Reports Announcements.

- Retrospective Search Service (August 1964 to date)
  - With Engineering Review $185
  - Without Review $ 90

- Retrospective and Current Awareness
  - With Engineering Review $275
  - Without Review $135

ENGINEERING INDEX COMPENDEX

- Custom Profiles

  - Current Awareness Service (12 consecutive months)
    - With Engineering Review $250
    - Without Review $200

  - Retrospective Search Service (1968* to date)
    - With Engineering Review $200
    - Without Review $150

  - Retrospective and Current Awareness
    - With Engineering Review $400
    - Without Review $300

*For profiles not relevant to the subject fields of plastics or electrical engineering retrospective search service will cover the period 1970 to date.

COSMIC

Documented computer programs developed by or for the National Aeronautics and Space Administration and the Department of Defense which have been offered for sale through the Computer Software Management and Information Center (COSMIC) are available through the

KASC. These computer programs are available in tape or card form, and their related documentations are available separately. Program descriptions on request from KASC.

DOCUMENT SERVICES

Full copy of most documents announced by KASC search services are available through the KASC at the following rates:

- Hard copy reproduced by KASC - $0.05/page of original document
- Duplicate microfiche - $0.50/per sheet
- Hard copy or microfiche obtained outside KASC - At Cost

A $0.50/document handling charge and postage costs are in addition to the fees determined by the above rates.

CATEGORICAL DESCRIPTION OF

STANFORD UNIVERSITY INFORMATION SYSTEM

Prepared for: Office of Science Information Services
National Science Foundation

Prepared by: Systems Development Division
Institute for Computer Sciences and Technology
National Bureau of Standards

January 1973

CONTENTS

CATEGORICAL DESCRIPTION OF

STANFORD UNIVERSITY INFORMATION SYSTEM

1. GENERAL DESCRIPTION

1.1 Project and Organization Name:

Stanford Public Information Retrieval System (SPIRES)
Institute for Communication Research
Stanford University
Stanford, California

1.2 Objectives and Operational Philosophy:

SPIRES has two long-range goals. The first is to provide a user-oriented, interactive, production on-line information storage and retrieval system for a variety of research groups in the Stanford community. The second is to support the automation efforts of University libraries (Project BALLOTS) by contributing to common software development. An immediate short-range goal has been to provide an on-line bibliographic information service for Stanford physicists, particularly for high-energy physicists.

1.3 Historical Background:

In 1967, a comprehensive user study was conducted on a target population of physicists. This study established information needs and priorities as a basis for system design. In late 1967, a small, one-terminal demonstration system was installed on the 360 model 75 computer at the Stanford Linear Accelerator Center using an IBM 2250 display terminal. Following the demonstration of the pilot system, most of 1968 was spent in creating the software necessary for a multiple-user on-line system. This included the development of an on-line supervisor program, and of search, retrieval, and update programs. By early 1969, SPIRES I had been tested and was ready for service; in late February operation began for an hour and a half a day, five days a week. This service schedule continued through the summer of 1969. After several months of operational experience, the last quarter of 1969 was spent in evaluating the SPIRES I system. The experience with SPIRES I was the basis for a six-phase development cycle defined for a SPIRES II production system. SPIRES I has continued in operational service at

SLAC since 1970. The first phase of the SPIRES II system development process was completed during the first quarter of 1970. This first phase of preliminary analysis characterized the users and the user environment and summarized the limitation of SPIRES I. It then went on to outline a long-range scope of retrieval and file management capabilities as well as the first implementation of SPIRES II. The second phase was well underway in July 1970. System requirements were established and approved by project staff and system users. A variety of technical tasks were carried out: the evaluation of existing programming languages and software, system simulation, the writing of an on-line command language, the designing of an analyzer to parse the language.

1.4 Present Operational Status:

SPIRES I is self-supporting and has been in operational service at SLAC since 1970. It is accessible on the IBM 360/91 from any SLAC terminal and from terminals on campus and elsewhere that have telephone dialup and entree to a SLAC account. The data base consists of 18,000 high energy physics preprints. Preprints in Particles and Fields, a publication based on the weekly input to the preprint file, has 600 subscribers, not including those at SLAC. A DESY high energy physics file is being built; it now contains approximately 45,000 entries.

SPIRES II basic software is coded and tested. The system was formally released to the public on October 19, 1972.

1.5 Future Plans:

SPIRES staff expects to complete work in seven areas:

*(1) File definition processor
*(2) The processing rules
*(3) The On-Line Semantic Modules
**(4) Documentation
***(5) Software Optimization
***(6) Statistics Gathering
*(7) Policy and Procedures for Operating the System

* Reported complete on 10/20/72
** Due for completion 12/31/72
*** Continuing tasks dependent on a period of operational services.

2. ADMINISTRATIVE DETAILS

2.1 Principal Investigator:

David Phillips
Director, Campus Facility of Stanford Computation Center
Cypress Hall, Stanford University
Stanford, California 94305
(415) 321-2300, Ext. 2755

2.2 User Community:

The primary user groups of SPIRES I are physicists and the library staff at SLAC.

The largest user group of SPIRES II is the Stanford library. An 18,000 record MARC file is operational for their use (and for searching by anyone in the Stanford community). Two graduate classes are using SPIRES for course notes, papers, and bibliographies.

2.3 Documentation:

(1) SPIRES (Stanford Public Information Retrieval System) 1970-1971. Annual Report to NSF, OSIS, December 1971.

(2) Design of the Stanford Public Information Retrieval System, July 1971.

(3) Draft copy of Volume I "Introduction to the SPIRES System."

(4) Draft copy of Volume II "How to Select and Search a File."

(5) Draft copy of Volume III "How to Build or Modify a File."

(6) Annual Report of Research in Progress, Institute for Communication Research, Stanford University, February 1972.

(7) SPIRES (Stanford Public Information Retrieval System) 1968. Annual Report to NSF, OSIS - January 1969.

(8) Campus Computer Facility, Services and System Overview, Stanford University, April 1972.

2.4 Services and Costs:

Type of services and costs are still in the process of being defined. Some cost analysis studies

show cost as consisting of three components:

(1) Storage Cost

(2) Build Cost

(3) Search Cost

(1) Storage Costs - Stanford Computation Center is presently charging disk rate with two plans:

Plan (A) rent a disk pack for $800 per spindle per month.
Plan (B) rent storage on a block (2048 byte) for $.007 per day.

Storage cost is the biggest cost to the user. By the end of the year, dismountable disk packs capable of load/unload to tapes, should be available, and these should reduce storage costs.

(2) Build Cost - This cost is a function of the number of indexed terms per record. (For example, for the MARC file, it is currently costing 15 to 20 cents per data record; for PLANTBIO, it is costing approximately 7 cents per data record; for the PROSTATE file, it is costing 50 cents per data record.) These costs include overhead and consulting, but not system development.

(3) Search Cost - Search cost is independent of file size but is dependent on the complexity of the search command. For a typical command which takes 63 ms. processing time, search cost is less than 3 cents. The average search cost is approximately $12 per hour.

3. DATA BASES

3.1 Data Base Selection

3.1.1 Requirements and Acquisition Strategy: Stanford is interested both in serving individuals with small private files and in providing access to the large commercially available data bases. With the

library as the first major large user (the MARC file), and a number of users of "personal files," they feel they have both the dollar volume and variety of user support to keep the system viable. On-line disk storage costs constitute the chief problem for operational service with large files. Operational on-line systems have different cost problems than batch SDI systems. They may need to develop a networking capability to generate a sufficient number of users to amortize high storage costs for large files.

3.1.2 Discipline Coverage
Stanford does not have any specific discipline coverage. The PREPRINTS file on SPIRES I was carried over. There is a student file containing registration record information and there are several files being built in the area of medical application such as records on cancer patients and transplant data, etc.

3.2 Data Preparation, Conversion and Entry:
The user first establishes his file structure by interacting with the SPIRES II file definition mode. The user defines the types of records that will be in a file and also provides the rules which dictate how the elements of the file are to be processed and how they relate to each other. Data are then entered on-line as SPIRES "ADD" command. Two modes are available: the first technique involves adding new records on-line followed by the physical file updating overnight; the second method allows the addition of many records at one time during a batch run.

SPIRES allows flexible record structure ranging from singular, fixed occurence, fixed length format to multiple-variable length format. (See Volume IV of "How to Build or Modify a File" for detail on system commands.)

3.3 Data Base Contents

3.3.1 Data Base (1): MARC File
This is the first module of BALLOTS (Bibliographic Automation of Large Libraries Using a Time Sharing System) data base. It presently consists of 18,000

records and is expected to grow to 40K or 50K. Data entry is via tapes from the Library of Congress. It is searchable on author, title, LC number, etc.

3.3.2 Data Base (2): STUDENT
This is the registration record file for the College of Notre-Dame in California. It now has 300 records. All 25 data fields are inverted and are all searchable. It has a capability of 10K records, but access would be very slow with that large a data base.

3.3.3 Data Base (3): PREPRINT
This is a subset of the PREPRINTS file accessible through SPIRES I. It has 300 records.

3.3.4 Data Base (4): PROSTATE
This is the file of 64 elements of codified and numeric data on cancer patients. It also has 300 records.

3.3.5 Data Base (5): TRANSPLANT (Planned data base). Similar to PROSTATE containing data on transplants from Dr. Schumway's research.

3.3.6 Data Base (6): PLANTBIO (Planned data base). It consists of bibliographic data on photosynthesis. Size will be approximately 4K. This file is now in the process of being built.

3.3.7 Data Base (7): NUCLEAR POWER (Planned data base). It consists of environmental data for the Sloan Foundation. Size will be approximately 4K.

4. HARDWARE CONFIGURATION

4.1 Main Frame:
IBM 360/67 with PDP-11 "front end."

4.2 Core Size:
SPIRES II executes under ORVYL (Stanford Time-Sharing Monitor). The Monitor resides in 120,000 bytes of the partition. The remainder is divided into thirty-five 4,096-byte pages. SPIRES

executes as a subprocessor in the virtual memory. Total core size is 28 million bytes.

4.3 Mass Storage Devices:
Control Data double density disks.

4.4 Input Devices
Up to 90 typewriter (IBM 2741 or Datel type) and teletype terminals can be supported concurrently in addition to the CRT terminals supported via the PDP-11 "front-end" machine.

4.5 Output Devices:
Same as input plus 1403 printer, 2540P punch.

5. SOFTWARE CONFIGURATION

5.1 Operating System:
SPIRES II executes under ORVYL, the Stanford Time-sharing submonitor under OS/360 MFT-II.

5.2 Operation Environment:
The Stanford campus computing facility has an IBM 360/67 with OS/360 MFT-II. Some of the major partitions are: High-speed Batch consists of ALGOL W language processor; Large Batch consists of language processors such as FORTRAN, LISP, PL/1, PL360, COBOL, etc; ORVYL, the time-sharing submonitor which SPIRES II will execute; WYLBUR which is a text editor; MILTEN which is the terminal communications processor, and HASP which provides spooling of the unit record input and output to and from the batch partitions. ORVYL utilizes the time-sharing hardware peculiar to the model 67 called "dynamic address translation". ORVYL also has its own macro-like service code for doing input and output. SPIRES II executes as a subprocessor and is re-entrant. A user attached to SPIRES may enter WYLBUR text editor and hence edit and update his file with WYLBUR commands.

5.3 Information System:

5.3.1 Name and Brief Description:
Stanford Public Information Retrieval System (SPIRES) is a generalized data management system capable of handling both bibliographic data and numeric and structured data.

Spires II is heavily embedded in ORVYL. It consists of four components:

(1) On-line Module containing functions such as search, update, tutorial preparation and text-editing.

(2) Batch Update Module containing functions such as add data, delete data and change data in the batch mode.

(3) Utility Processor containing a master terminal language processor, statisical analysis routines, communication to peripherals and subprocessors.

(4) File definition Compiler containing syntax parser for data definition and capable of accepting data on-line.

The architecture of all four modules is the same consisting of approximately 72K of reentrant code. About 3% of total code is supervisor interface. The rest consists of a parser, a set of action lists, a set of semantic interpretation rules and file service routines containing SPIRES virtual access method.

5.3.2 Source Language:
90% PL/1 for SPIRES I.
100% PL360 for SPIRES II. (PL360 is a language with machine-level functions with ALGOL-like features.)

5.3.3 Mode:
Mode for searches is on-line or batch.
Mode for file updates is pseudo on-line. The updates are entered via a terminal but placed in a deferred queue and updated overnight.

5.3.4 Generalized Packages Used:
SPIRES II can communicate to WYLBUR which is a text editor.

The structure of SPIRES II itself is table-driven and hence consists of a generalized parse program.

5.3.5 Availability:
No administrative procedure has been worked out for software distribution.

6. COMPUTER PROCESSING FUNCTIONS

6.1 Data Definition:
There is a generalized data definition facility whereby a user may define his record structure interactively. Detailed instructions on how to define a file will be documented in Vol. IV of "Design of the SPIRES II".

6.2 Data Maintenance:
Data Maintenance can be performed both as batch update or as on-line entry followed by physical file writing overnight. The Maintenance language is very flexible, and a user may enter from SPIRES to a working data set in WYLBUR and hence modify any data using standard WYLBUR text editing commands. The detailed instructions on how to modify a file are documented in Vol. III of "Design of the SPIRES II".

6.3 Data Retrieval:
A record in SPIRES must contain the following three components: The key data element, the point group, and values of data elements. These records physically reside in the blocks which are 2,048 bytes in length. The records in the data set are logically arranged in either a series of simple, fixed-length slots or in a tree structure. Retrieval for the fixed-length slot structure is straight forward and quick. An algorithm for searching records organized in a tree-structure is explained on p. 4-3 of "SPIRES 1970-1971 Annual Report to NSF".

The search that a user of the SPIRES performs involves an iterative process which starts with a basic criterion and is then restricted or enlarged by the introduction of additional criteria. The logical operators allowed are "AND", "AND NOT" "OR" and "NOT". The qualifier operators consist of words such as "BEFORE", "AFTER", "WITH", "STRING", "BETWEEN", "MASK", etc.. The search language is user-oriented and explained in Vol. III of "Design of SPIRES II".

6.4 Data Output:
The system is not geared to massive output. The output consists of response to queries. The user also may put the data to be output in a WYLBUR working set and thus use WYLBUR commands to either print out complete records or list certain desired segments of the goal records. There exists a PAGE and SCROLL command to be used when CRT video terminals are operational.

6.5 Special:

(1) SPIRES has THESAURUS and SYNONYM capability built in if thesauri and synonyms have been defined and constructed for a file.

(2) Some disgnostic features have been built in SPIRES with commands called EXPLAIN and HELP. If during the search process the user becomes confused or misdirected, he may type "HELP" and the system will provide information about possible commands that may be applicable at that particular point.

(3) There is a BACKUP command allowing the user to see the search result from the previous iteration.

(4) There are class privileges for users in accessing and updating certain files.

(5) The SPIRES II file design has provision for quick recovery and programmatic validation by redundant storage of critical control information.

7. USER INTERFACE

7.1 Current Awareness Service Available? No.

7.2 Batch Retrospective Search Available? No.

7.3 Interactive Retrospective Search Available? Yes.

7.3.1 Query Language:
The query language consists of two main commands: SELECT and FIND. SELECT designates the files the searcher is interested in, and FIND, followed by the appropriate search term causes the system to gather a list of records that conform to the criteria stated as part of the command.

By applying logical and qualifier operators, the user can cause the FIND command to perform complex functions. The query language is described in Vol. II of "Design of SPIRES II."

7.3.2 Training and Assistance:
Formal classes are offered by the User Services Group of the Stanford Computation Center. First classes began in October 1972.

7.3.3 Response Time:
No calculations on a typical response time is quoted. We were told that the search time is independent of file size but is dependent on the complexity of the search command. A typical search command may take 63 msec processing time.

7.3.4 Charges:
See Section 2.4

8. TRANSFERABILITY CHARACTERISTICS AND EXPERIENCES

8.1 Software Transferability:
The SPIRES II software is heavily embedded in ORVYL monitor and SPIRES II is coded in PL360. Hence transferring SPIRES involves transferring a PL360 translator and the ORVYL time-sharing monitor.

8.2 Data Base Transferability:
The major files in SPIRES II are private files.

8.3 Methodology Transferability:
The implementation of SPIRES employs advanced concepts in language parsing techniques. The SPIRES II command language is defined in Action BNF and there is a generalized parser which performs the action guided by the BNF definition with the execution of the semantics routine. The concept is generalized enough that one can build a language processor or data management access package with the same parsing routine.

The tree-balancing algorithm to permit fast access to the search processing is a methodology worth noting. The algorithm is described on p. 4-5 of "SPIRES 1970-1971 Annual Report to NSF."

☆ U.S. GOVERNMENT PRINTING OFFICE: 1973--511-326/270

FORM NBS-114A (1-71)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET	1. PUBLICATION OR REPORT NO. NBS TN-781	2. Gov't Accession No.	3. Recipient's Accession No.

4. TITLE AND SUBTITLE

A Study of Six University-Based Information Systems

5. Publication Date: June 1973

6. Performing Organization Code

7. AUTHOR(S)
B. Marron, E. Fong, D. W. Fife and K. Rankin

8. Performing Organization

9. PERFORMING ORGANIZATION NAME AND ADDRESS

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20234

10. Project/Task/Work Unit No.
640-4404

11. Contract/Grant No.
Interagency Agreement #NSF - CA68

12. Sponsoring Organization Name and Address

National Science Foundation
18th & G. St. N.W.
Washington, D. C. 20550

13. Type of Report & Period Covered
Interim

14. Sponsoring Agency Code

15. SUPPLEMENTARY NOTES

16. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)

A methodology for categorically describing computer-based information systems was developed and applied to six university-based, NSF-supported, systems. The Systems under study all operate as retail information centers primarily serving campus communities by accessing large commercially-available data bases using 3rd generation computer configurations. The systems vary in design philosophy, mode of user service, transferability characteristics, and operational status. A summary matrix is included.

17. KEY WORDS (Alphabetical order, separated by semicolons)
Computer-based systems; information systems, university; university computer systems.

18. AVAILABILITY STATEMENT

[X] UNLIMITED.

[ ] FOR OFFICIAL DISTRIBUTION. DO NOT RELEASE TO NTIS.

19. SECURITY CLASS (THIS REPORT)
UNCLASSIFIED

21. NO. OF PAGES
98

20. SECURITY CLASS (THIS PAGE)
UNCLASSIFIED

22. Price
$1.25 Domestic Postpaid; $1.00 G.O.P. Bookstore

USCOMM-DC 66244-P71

NBS TECHNICAL PUBLICATIONS

PERIODICALS

JOURNAL OF RESEARCH reports National Bureau of Standards research and development in physics, mathematics, and chemistry. Comprehensive scientific papers give complete details of the work, including laboratory data, experimental procedures, and theoretical and mathematical analyses. Illustrated with photographs, drawings, and charts. Includes listings of other NBS papers as issued.

Published in two sections, available separately:

• Physics and Chemistry (Section A)

Papers of interest primarily to scientists working in these fields. This section covers a broad range of physical and chemical research, with major emphasis on standards of physical measurement, fundamental constants, and properties of matter. Issued six times a year. Annual subscription: Domestic, $17.00; Foreign, $21.25.

• Mathematical Sciences (Section B)

Studies and compilations designed mainly for the mathematician and theoretical physicist. Topics in mathematical statistics, theory of experiment design, numerical analysis, theoretical physics and chemistry, logical design and programming of computers and computer systems. Short numerical tables. Issued quarterly. Annual subscription: Domestic, $9.00; Foreign, $11.25.

TECHNICAL NEWS BULLETIN

The best single source of information concerning the Bureau's measurement, research, developmental, cooperative, and publication activities, this monthly publication is designed for the industry-oriented individual whose daily work involves intimate contact with science and technology—*for engineers, chemists, physicists, research managers, product-development managers, and company executives.* Includes listing of all NBS papers as issued. Annual subscription: Domestic, $6.50; Foreign, $8.25.

NONPERIODICALS

Applied Mathematics Series. Mathematical tables, manuals, and studies.

Building Science Series. Research results, test methods, and performance criteria of building materials, components, systems, and structures.

Handbooks. Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Special Publications. Proceedings of NBS conferences, bibliographies, annual reports, wall charts, pamphlets, etc.

Monographs. Major contributions to the technical literature on various subjects related to the Bureau's scientific and technical activities.

National Standard Reference Data Series. NSRDS provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated.

Product Standards. Provide requirements for sizes, types, quality, and methods for testing various industrial products. These standards are developed cooperatively with interested Government and industry groups and provide the basis for common understanding of product characteristics for both buyers and sellers. Their use is voluntary.

Technical Notes. This series consists of communications and reports (covering both other-agency and NBS-sponsored work) of limited or transitory interest.

Federal Information Processing Standards Publications. This series is the official publication within the Federal Government for information on standards adopted and promulgated under the Public Law 89–306, and Bureau of the Budget Circular A–86 entitled, Standardization of Data Elements and Codes in Data Systems.

Consumer Information Series. Practical information, based on NBS research and experience, covering areas of interest to the consumer. Easily understandable language and illustrations provide useful background knowledge for shopping in today's technological marketplace.

BIBLIOGRAPHIC SUBSCRIPTION SERVICES

The following current-awareness and literature-survey bibliographies are issued periodically by the Bureau:

Cryogenic Data Center Current Awareness Service (Publications and Reports of Interest in Cryogenics). A literature survey issued weekly. Annual subscription: Domestic, $20.00; foreign, $25.00.

Liquefied Natural Gas. A literature survey issued quarterly. Annual subscription: $20.00.

Superconducting Devices and Materials. A literature survey issued quarterly. Annual subscription: $20.00. Send subscription orders and remittances for the preceding bibliographic services to the U.S. Department of Commerce, National Technical Information Service, Springfield, Va. 22151.

Electromagnetic Metrology Current Awareness Service (Abstracts of Selected Articles on Measurement Techniques and Standards of Electromagnetic Quantities from D-C to Millimeter-Wave Frequencies). Issued monthly. Annual subscription: $100.00 (Special rates for multi-subscriptions). Send subscription order and remittance to the Electromagnetic Metrology Information Center, Electromagnetics Division, National Bureau of Standards, Boulder, Colo. 80302.

Order NBS publications (except Bibliographic Subscription Services) from: Superintendent of Documents, Government Printing Office, Washington, D.C. 20402.

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards
Washington, D.C. 20234

OFFICIAL BUSINESS

Penalty for Private Use, $300

POSTAGE AND FEES PAID
U.S. DEPARTMENT OF COMMERCE
COM-215