# Lawrence Berkeley National Laboratory
## Recent Work

**Title**

SYSTEM OF INFORMATION RETRIEVAL AND ANALYSIS FOR PLANNERS (SIRAP)

**Permalink**

https://escholarship.org/uc/item/1q88k98q

**Author**

Gey, Fredric.

**Publication Date**

1974-04-01

# SYSTEM OF INFORMATION RETRIEVAL AND ANALYSIS FOR PLANNERS (SIRAP)

Fredric Gey

April 1974

# DISCLAIMER

System of Information Retrieval and Analysis
For Planners (SIRAP)


Fredric Gey
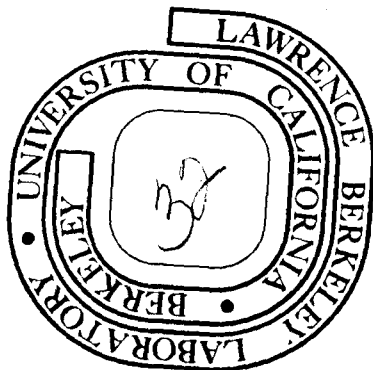

Mathematics and Computing Group
Lawrence Berkeley Laboratory
University of California
Berkeley, California

## ABSTRACT

SIRAP is a system of information retrieval and analysis for government planners. SIRAP consists of three distinct, somewhat overlapping entities, the SIRAP Information Retrieval Facility, the SIRAP Data Analysis Facility and the SIRAP Teletype Facility.

All data files within SIRAP are fixed statistical, economic, environmental, and demographic data obtained from U.S. Government agencies such as the Bureau of Census, Economic Development Agency, Bureau of Labor Statistics, and the U.S. Army Corps of Engineers. Most SIRAP files are designed to be uniformly retrievable using the QWICK QWERY retrieval program (a proprietary product of Consolidated Analysis Centers, Inc.). Some large files (such as the 1970 U.S. Census of Population) have their own random access retrieval system coupled with fixed format computer reports.

## INTRODUCTION

Traditional approaches to planning for governmental projects have suffered from inadequate information and analytical facilities. This has led to decisions based upon insufficient information, upon excessive political considerations, and, not uncommonly, on pure guesswork. As the budget of the U.S. Government approaches 1/3 of the gross national product in the United States, federal agency spending can have vast primary and secondary impacts on regional and local economies. Thus it becomes important that such funds be allocated wisely, based upon the most complete available information for planning decisions.

The Planning Division of the U.S. Army Corps of Engineers has, for more than thirty years, been required to file cost/benefit analysis reports upon all civil works projects that agency proposes to undertake. Until recently, information for such reports has been based upon individual library research by planners, or upon subcontracted literature search and information extraction. About two years ago the Corps decided to devise an information system to meet planning needs. Because of favorable cost considerations and certain unique hardware capabilities, the Lawrence Berkeley Laboratory (LBL) Computing Facility was chosen as the resident site for this information retrieval and analysis system, known as SIRAP (System of Information Retrieval and Analysis for Planners). Development of this system at LBL has beeen carried out by members of the Research and Development staff of the Mathematics and Computing Group at LBL. Development of the data base in a uniform manner has been carried out by Consolidated Analysis Centers, Inc. under contract for the Office of Chief of Engineers, U.S. Army Corps of Engineers, Washington.

Portions of SIRAP have been duplicated and distributed regionally to Army Corps Division office computing networks. This paper will not discuss this aspect of SIRAP, but will focus upon development at the Lawrence Berkeley Laboratory.

## EXTERNAL ACCESS TO SIRAP FACILITIES

Figure 1 represents a schematic diagram of the Lawrence Berkeley Laboratory Computing Center. The central computer is a CDC-7600. This is front-ended by a CDC-6600 – CDC-6400 computing network to handle I/O, batch job submission, and remote teletype access. Remote batch communication is handled, by two COPE processors[1]. The typical SIRAP user will submit a computer job either locally or via a remote batch station, which will then be controlled by the front-end computers, entered to and executed by the CDC-7600 (with tape and other mass storage access controlled by the 6000s) with results returned to the user at job conclusion

either by on-site printing (including facilities for computer output on microfilm) or queued disk file for remote batch stations. Certain SIRAP data files and analytical programs are available via teletype. This is accomplished by dial-up (or hard-wired cable) to the RECC (Remote Equipment Control Computer] system, a PDP-8 hardware-software system for teletype control and communication with the CDC-6600 computer. RECC supports up to 125 teletype ports. Software support on the CDC-6600 computer is provided under the SESAME interactive system[2].

## SIRAP INTERNAL HARDWARE CONFIGURATION

SIRAP data which is uniformly retrievable via QWICK QWERY is stored in the LBL computer tape library. Each tape is assigned a unique identification number and SIRAP writeups have a tape table cross tabulating logical file names (which may be multi-reel) with their physical equivalent. Any tape may be accessed by punching up a single control card.

Two large files, the fourth count of the 1970 U.S. Census of Population (approximately equivalent to 400 reels of tape) and County Business Patterns (for 1967, 1968, 1970 and 1972 at the four digit U.S. Standard Industrial Classification Code level) are stored on a random access mass storage device, the IBM-1360 Photo Digital Storage System (or Chipstore, for short). All SIRAP retrieval and analysis programs, as well as QWICK QWERY data dictionaries, are stored on an IBM-2321 Data Cell. To retrieve SIRAP data using QWICK QWERY, only two cards need to be punched up and added to a fixed set of control cards (one specifying the correct data tape, the other specifying the dictionary which is a directory to the data on that tape) in addition to the QWICK QWERY data deck. Figure 2 is a schematic diagram of the SIRAP internal hardware configuration.

## SIRAP LOGICAL STRUCTURE

The logical structure of SIRAP is deliniated in Figure 3. SIRAP consists of three distinct, somewhat

overlapping facilities: the SIRAP Information Retrieval Facility, the SIRAP Data Analysis Facility, and the SIRAP Teletype Facility. SIRAP is structured in this way because it is not a computer *system* as such, i.e. SIRAP does not have a uniform data structure imposed on all data, with uniform input/output capabilities for participant programs, and with all programs put into operation via a single executive system. With the limited funding available, the decision was made to purchase and install the best among available standard software packages, and to pursue software development only where the software available was inadequate to the specialized needs of SIRAP. This was particularly the case with the SIRAP Teletype Facility.

Despite these drawbacks, from a government planning point of view, SIRAP is a system. It provides inexpensive computer access to a vast amount of planning data (now nearly 300,000,000 different data items) together with the analytical tools necessary to utilize this data for planning purposes.

## SIRAP INFORMATION RETRIEVAL FACILITY

The *SIRAP Information Retrieval Facility* consists of the SIRAP data base and those programs for retrieval of SIRAP data. The salient feature of SIRAP is that much of its data is uniformly retrievable using the QWICK QWERY retrieval and report generation program. The following data files are uniformly retrievable using QWICK QWERY.

1970 U.S. Census of Population and Housing, 1st and 4th count (data by county, census tract and enumeration district/block group)

1970 Domestic and International Transportation of U.S. Foreign Trade

1950-2020 OBERS Projections of U.S. Economic, agricultural and employment activity (data by economic planning area, water resource planning area, state and SMSA)

1949, 1959, 1964 U.S. Census of Agriculture (data by county)

1963, 1967 Location of Manufacturers in the U.S. (4 digit SIC level, by county)

1960 U.S. Census of Population (by county)

1952, 1960, 1962 County Merge (1960 Census of Population,1952, 1962 City-County Data Book, Bureau of Public Roads data merge)

INCOME (Bureau of Economic Analysis income data, 1929-1969, by county).

This data is stored as fixed-length, blocked binary records of character strings. QWICK QWERY handles all character manipulation for data extraction operating in sequential mode only. Record sizes vary from 96 characters (for the 1963 Location of Manufacturers File) to 11960 characters (for the Bureau of Economic Analysis Income File). All files interface to QWICK QWERY through a single read subroutine integrated into the program.

QWICK QWERY offers the SIRAP user the following capabilities:

1. Selective retrieval and filtering
2. Three levels of sorting
3. Totals and subtotals
4. Indexing and new data generation based on straightforward or conditional computation
5. Standard or custom (user-generated) report generation
6. BCD and binary output file options

This last is of special significance since BCD ouput files from QWICK QWERY may be read by SPSS (Statistical Package for the Social Sciences) for further analysis.

1970 U.S. CENSUS OF POPULATION

Each record of the 1970 U.S. Census of Population (Fourth Count) consists of 1178 census data items tabulated for 5 racial/ethnic breakdowns (total, white, black, Spanish American, American Indian, and other). This data is available for the following geographic areas:

State
Standard Metropolitan Statistical Area
Urbanized Area
County
Minor Civil Division/Census County Division
Place (2500 or greater population)
Census Tract

These 78,000 records of data have been stored on an IBM Data Chip Storage device under an index sequential retrieval system developed by the Socio-Demographic Data Base project at LBL. Retrieval is available to SIRAP users in the form of two fixed-format reports generated by the project for planners of the U.S. Department of Labor, Manpower Administration. The first, *Summary Manpower Indicators*, is a 14-table summary of population characteristics prefaced by a two-page written summary. The second, *Detailed Manpower Indicators*, is a 90-table, 136-page report which exhaustively describes the population characteristics of a census geographic area. Since, under the system, geoarea records may be *aggregated*, other areas (such as water basins) may be approximated by census records and a report issued on the aggregate record.

A file on County Business Patterns (from the Bureau of Economic Analysis, U.S. Department of Commerce) has been installed on the chipstore for SIRAP retrieval. Records are at the state and county level for four digit U.S. Standard Industrial Classification Codes[4] for the years 1967, 1968, 1970 and 1972. This data was not implemented for QWICK QWERY retrieval because it is in variable length records, with numerous omissions due to disclosure/suppression to protect individual privacy.

SIRAP DATA ANALYSIS FACILITY

The *SIRAP Data Analysis Facility* contains two major statistical analysis programs, GRASS (Generalized Research Analysis Statistical System)[5] and SPSS (Statistical Package for the Social Sciences)[6]. SPSS operates in batch mode only, while GRASS operates in both batch and remote teletype mode. The two have similar capabilities however

because SPSS is the more powerful, it is our recommended package for SIRAP users in batch mode.

SPSS offers the SIRAP user the following tabulation and statistical computation capabilities:

1. Frequency distributions and histograms
2. Cross-Tabulations
3. Basic statistics
4. Nonlinear regression
5. Pierson correlations
6. Factor analysis
7. Univariate analysis of variance
8. Social science oriented statistical computations

SPSS can operate directly on BCD records output by the QWICK QWERY retrieval program.

## SIRAP TELETYPE FACILITY

The LBL Computer Facility is a batch-oriented center with limited support of interactive computing through systems on the CDC-6600 computer. With this in mind, the major thrust of SIRAP has been toward the development or installation of large programs operated in batch mode. However, as time and manpower permit, certain data files and programs are being made available to the SIRAP user via remote teletype. At this point, the OBERS Projections File and the County Business Patterns File have been implemented for teletype retrieval. These files have been indexed for direct retrieval on a geographic area code as the key. County Business Patterns is additionally indexed by year of data accumulation.

In addition to the implementation of GRASS for teletype usage, three small programs for linear regression and correlation coefficient calculation have been written specifically for SIRAP teletype usage.

## SIRAP DOCUMENTATION

Experience has demonstrated the futility of trying to maintain up-to-date versions of printed manuals documenting a computer system that is extensive, varied, and subject to continual modification. This problem was recognized several years ago by the Users Services Group

of Mathematics and Computing at the Lawrence Berkeley Laboratory. Their solution of the problem, which has been borrowed by the SIRAP project, is to implement a dynamic documents library, stored on the IBM-2321 Data Cell. To do this, two programs are necessary: a card-image update utility and a card-oriented text editing program. The former is supplied as the standard CDC software utility UPDATE. The latter has been fulfilled by a locally-written text editing program BARB.

SIRAP writeups are stored on the data cell in two forms: as a BARB input deck in UPDATE program library format and as a BARB output file which has been formatted, indexed, with table of contents, ready for computer retrieval.

In this way, additions and changes to the writeups can be made without the expense and delay of publishing, and the writeups are very accessable. Even at a remote site, the user can have the latest version of any SIRAP writeup printed out at any time.

SIRAP documentation consists, at present, of five separate writeups (about 90 pages in total)

Introduction to SIRAP (SRINTRO)
SIRAP Information Retrieval Facilities at LBL (SRINFO)
SIRAP Interactive Facilities at LBL (SRTTY)
SIRAP Data Analysis Facilities at LBL (SRANL)
Summary of the Socio-Economic and Environmental Data Base at LBL

These writeups may be obtained directly from the computer by running the following job:

JOBCARD
LIBCOPY, WTSIRAP, DUM/BR.
COPY, DUM, OUTPUT.
<END-OF-JOB-CARD>

## THE FUTURE OF SIRAP

SIRAP is one of five projects undertaken by the LBL Mathematics and Computing Group for various U.S. Government agencies. The others are:

Socio-Demographic Data Base Project

Computerized Mapmaking Project

Data Base Management Systems Project

Energy-Environmental Impact Modeling Project

The fruits of these specific, agency-oriented projects will be incorporated into SIRAP as time and funding permit.

SIRAP development is now one year old. During the next year, the major SIRAP developments are planned to be in the areas of data display (computerized map generation of SIRAP data) and data analysis (implementation of the Economic Statistic Package (ESP), the General Purpose Simulation System (GPSS), and the Input-Output modeling techniques of the Harvard-EDA Multi-Regional Input-Output model).

## ACKNOWLEDGMENTS

There is a tendency to assume that the author of a paper about a computing system is necessarily the man who created the system singlehandedly. This is certainly not the case of SIRAP. Gerry Litton implemented GRASS and the teletype analytical programs. Bill Hogan installed, improved, and maintains SPSS. Bob Healey designed and implemented the 1970 Census data base. Tricia Coffeen implemented the County Business Patterns File and aided in the installation of QWICK QWERY retrievable data base. I installed the QWICK QWERY retrievable data base, developed the SIRAP documentation, and generally handled all the grubby administrative details of the project.

I would like to thank Doris Nicholson who typed the paper, Ed Romascan who taught me how to edit the paper on the IRATE interactive editing system at LBL, and Jim Miller who printed the paper on his ARDIE computer typesetting system.

## REFERENCES

[1]    LBL, Users Services Group, BKY – User's Introduction to the Lawrence Berkeley Laboratory Computing Facility, UCID-3539, Feb. 16, 1974.

[2]    Greiman, W., SESAME, A New Interactive System, Mathematics and Computing internal document, Feb. 1974.

[3]    CACI, Inc., QWICK QWERY User's Manual, Consolidated Analysis Centers, Inc., 12011 San Vicente Blvd., Los Angeles, CA 90049.

[4]    1967 U.S. Standard Industrial Classification Code Manual, Executive Office of the President of the United States, Office of the Budget.

[5]    Cpt. Charles Messa, et.al. Generalized Research Analysis Statistical System, Lab. Report 329, U.S. Army Medical Research and Nutrition Laboratory, Fitzsimmons General Hospital, Denver, CO 80240.

[6]    Nie, N. H., et.al., SPSS Manual, New York, McGraw Hill, 1970. (Program distributed by Vogleback Computer Center, Northwestern University, Evanston, ILL.)
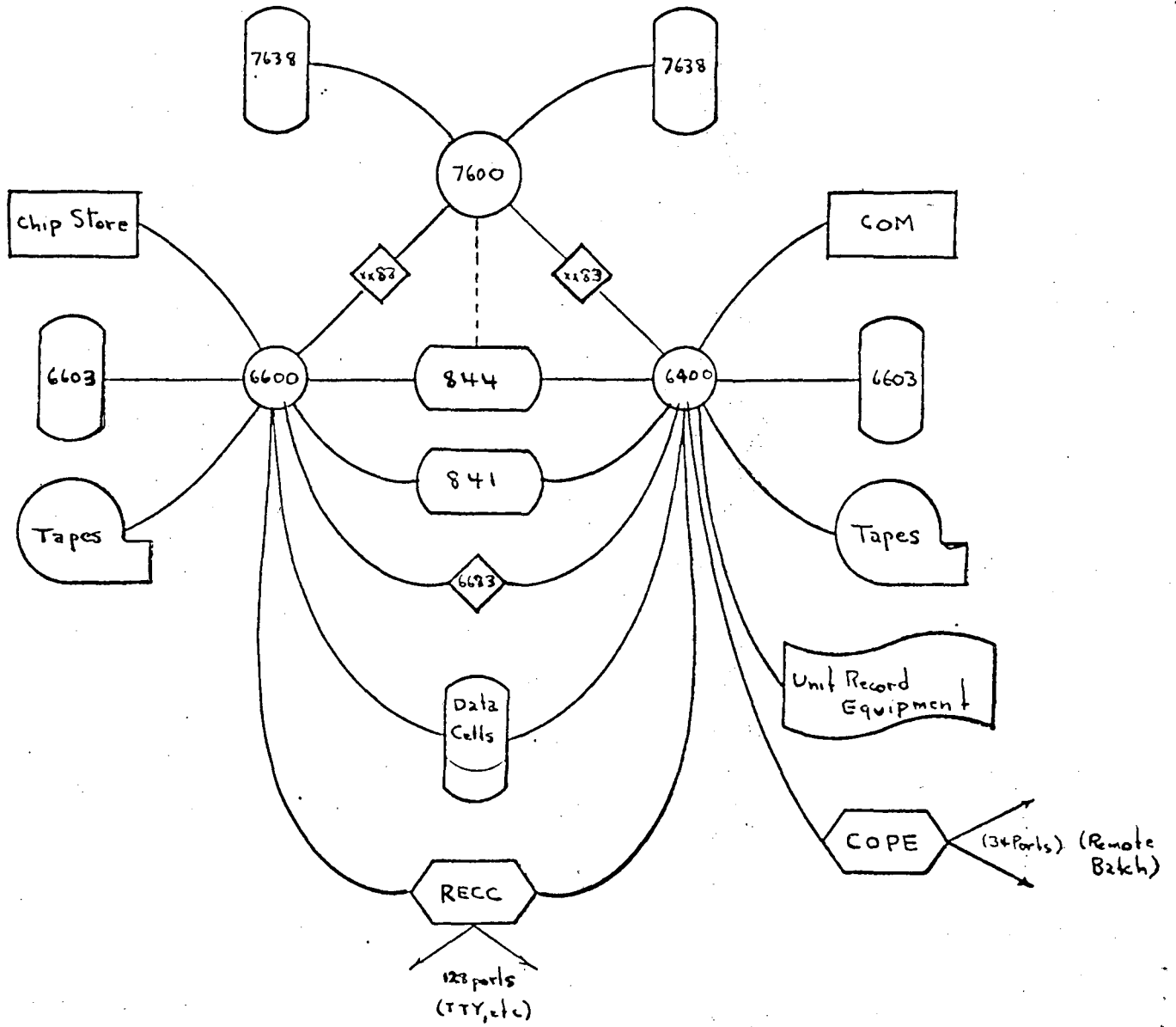
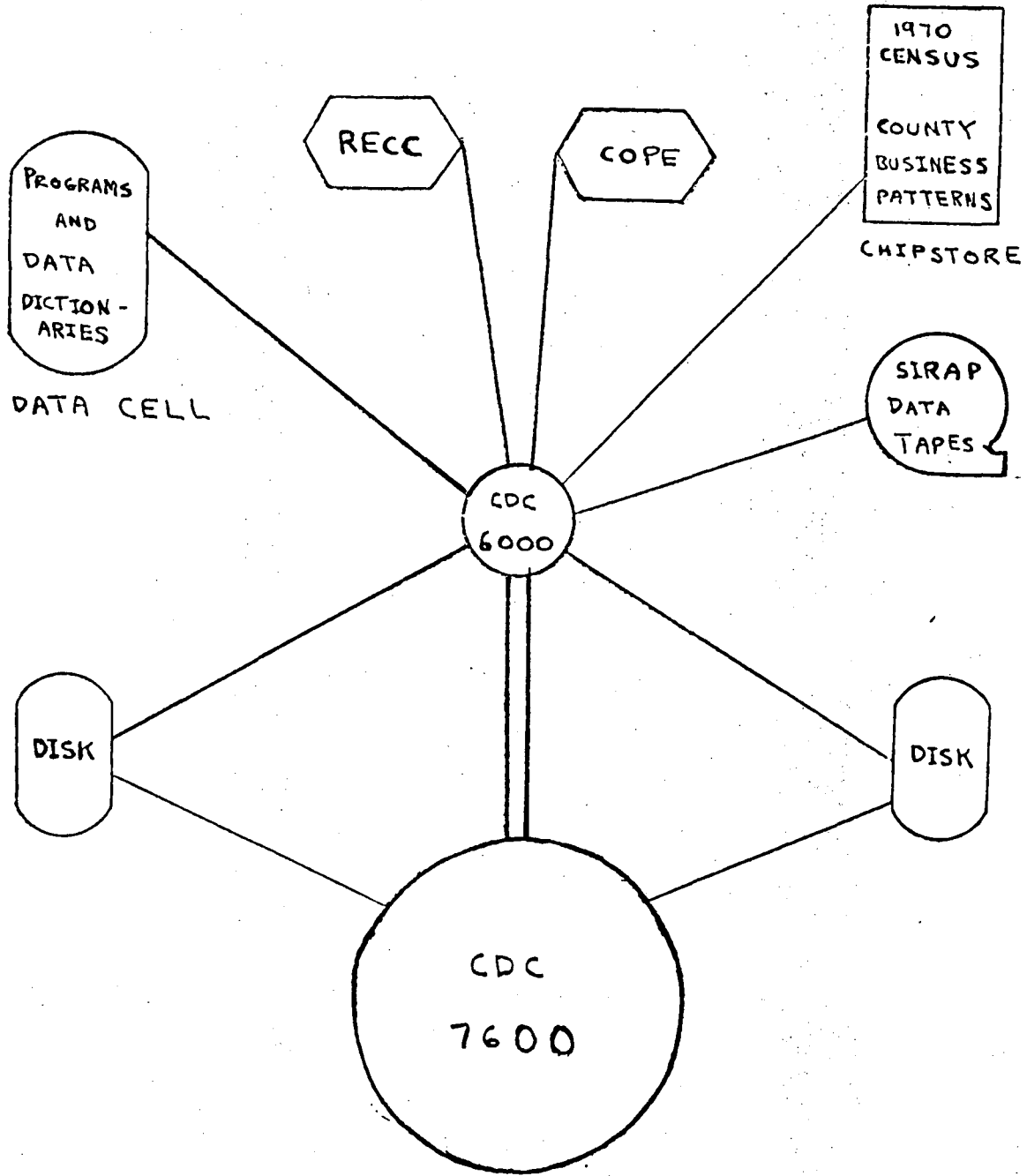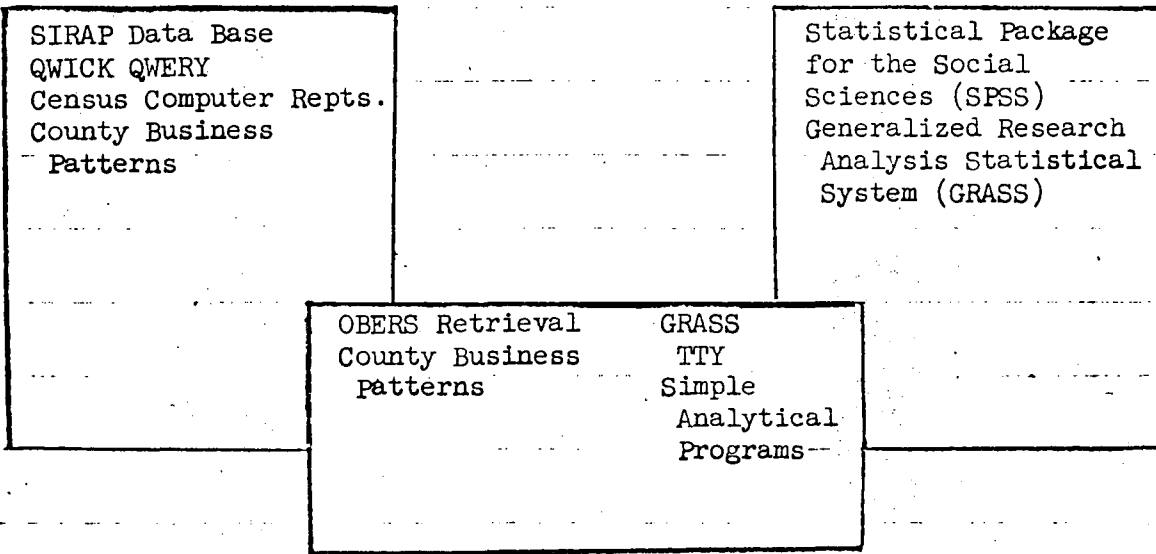LAWRENCE BERKELEY LABORATORY

COMPUTER CENTER



FIGURE 1

FIGURE 2

SIRAP Internal Hardware Configuration

SIRAP Information
Retrieval Facility

SIRAP Data Analysis
Facility

```
┌─────────────────────────────┐          ┌──────────────────────────┐
│ SIRAP Data Base             │          │ Statistical Package      │
│ QWICK QWERY                 │          │ for the Social           │
│ Census Computer Repts.      │          │ Sciences (SPSS)          │
│ County Business             │          │ Generalized Research     │
│  Patterns                   │          │   Analysis Statistical   │
│                             │          │ System (GRASS)           │
│          ┌──────────────────┴──────────┴──┐                       │
│          │ OBERS Retrieval     GRASS      │                       │
│          │ County Business     TTY        │                       │
│          │   Patterns          Simple     │                       │
│          │                      Analytical│                       │
│          │                      Programs  │                       │
└──────────┤                                ├───────────────────────┘
           └────────────────────────────────┘
```

SIRAP Teletype Facility


FIGURE  3

SIRAP Logical Structure