To be published in proceedings of
URISA 77 Urban and Regional Information
Systems Association Conference,
Kansas City, MO, August 7 - 11, 1977

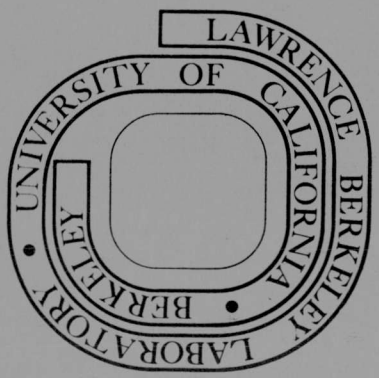
LBL-6417

THE REAP FAMILY OF COMPUTER PROGRAMS FOR RETRIEVAL OF SOCIO-ECONOMIC-ENVIRONMENTAL-DEMOGRAPHIC INFORMATION

Fredric Gey and Edna Williams

June 23, 1977

Prepared for the U. S. Department of Energy
under Contract W-7405-ENG-48



LBL-6417

Presented at URISA 77
Urban & Regional Information Systems Association
Conference, Kansas City, August 7-11, 1977

LBL - 6417

The REAP Family of Computer Programs

for Retrieval of

Socio-Economic-Environmental-Demographic Information

Fredric Gey and Edna Williams
Computer Science and Applied Mathematics Department
Lawrence Berkeley Laboratory
Berkeley, Ca 94720

ABSTRACT

REAP is a non-programmer-oriented computerized system for interactive retrieval of a large amount of socio-economic-environmental-demographic information. The REAP data base is mostly maintained at the county level for the entire United States; it has 11 files of information and allows retrieval at the record and data element level for approximately 10,000 named data elements. Thus the data base comprises about 30 million data items, and total size of 250-300 million characters.

This report was prepared for the U. S. Energy Research & Development Administration under Contract W-7405-ENG-48.

TABLE OF CONTENTS 29 JUL 77

1. INTRODUCTION

During the last five years the Lawrence Berkeley Laboratory's Computer Science and Applied Mathematics Group has designed a massive Socio-Economic-Environmental-Demographic Information System (SEEDIS) [1] of which REAP is a subsystem. Another such subsystem, the CARTE thematic mapping program was described in in URISA 76 [2].

REAP is intended to provide rapid access to information within SEEDIS where the user wants a limited amount of information (such as can easily be displayed on a teletype or CRT screen); it has been used by nationwide planning offices of the federal government (in particular, Department of Labor, Army Corps of Engineers, and Energy Research and Development Administration) for nearly two years now. The technical details of REAP were presented in [3].

Within the REAP family there are additional programs for batch retrieval of entire or partial records from the data base as well as a general report processor for producing reports complete with automated table of contents and index. The report processor is available within the SEEDIS monitor [4] in an 'online batch' mode of operation (one job control statement at a time) for online report debugging.

The next section of this paper will will describe the size of the data base and the information content of REAP files. Section 3 will discuss access to REAP within the SEEDIS monitor, which serves to mask the complexity of job control processing from the non-programmer user. Section 4 will describe REAP from the user viewpoint. Section 5 will briefly discuss the technical approach in both hardware and software utilized by the systems. The next sections will describe batch retrieval, data extraction and new file creation. Section 9 will describe future developments.

2. INFORMATION AVAILABLE VIA REAP

The information available within REAP is, in general, computerized data files made available by public data collection agencies of the United States government (primarily the Bureau of the Census and the Bureau of Labor Statistics, the Bureau of Economic Analysis, and the Economic Development Administration). Currently only one file (the Brookhaven National Labortory [5] county energy data base) contains derived data from various of these sources. Most data within REAP is maintained at the county level of geographic summary. Some files also contain state and SMSA data as well. The OBERS files (from the U. S. Water Resources Council) contain data for economic planning areas and water resource areas (defined as aggregates of counties).

The following files are currently installed and running under REAP -

File Name	Description	No. of Records	Size (Megabytes)
CCDBMRG	City County Data Book (1952-1972)	3,100	11.0
CNAGR49	1949 U.S. Census of Agriculture	3,100	12.6
CNAGR59	1959 U.S. Census of Agriculture	3,100	12.6
CNAGR64	1964 U.S. Census of Agriculture	3,100	12.6
CNAGR69	1969 + 1964 U.S. Census of Agriculture	3,100	23.7
CTY60PC	1960 U.S. Census of Population	3,100	30.7
OBERS	Series C OBERS Projections	695	4.5
OBERSE	Series E OBERS Projections	1,200	7.6
INCOME	BEA Income Time Series (1929-1969)	3,350	37.0
CNS4CT	Fourth Count Census Extract (1970)	3,400	12.6
ENERGY	Brookhaven County Energy Data Base	3,100	3.1

A general data base load feature is available, and new data files can be incorporated within a few days of their receipt at LBL.

3. ACCESS TO REAP - THE SEEDIS MONITOR

Most non-programmers approach computers with awe and unease; rightly so, for the complexities of job control language (JCL) specification can sometimes overwhelm even sophisticated programmers. In order to serve the access needs of information users as well as programmers, the SEEDIS monitor was created to provide a user-friendly interface to the complex data management, analysis, and display modules within SEEDIS.

The monitor performs many routine tasks by a single user command with a few options. It provides

- Straightforward access to SEEDIS modules
- Access to the file storage facilities at LBL
- Local computer mail service for SEEDIS users
- Uniform online help

In most cases the user can summon help at any point in the dialogue by typing "?" or "HELP", and instructions will be issued suggesting the appropriate input. Other major modules within SEEDIS include

FIFPAK - A 1970 Fifth Count Census Retrieval System
CHART - Graphic Display and Analysis System
CARTE - Thematic Mapping System
ESARS - Employment Applicant Data Retrieval System
LBLIRI - The LBL Interactive Resource Index

4. THE STYLE OF REAP

The major motivation behind REAP was to develop a friendly retrieval system which did not require the intervention of a programmer in the data retrieval process. This was done through a methodology of commands and prompting. Prompting by REAP carries the user gently through the three major steps of

file selection

study area definition

data display

The following global commands are available to the user to interrupt the usual chain of REAP execution at any time

command	meaning
HELP	provide help messages
LIST	list user input to current task
DISCARD	restart current task, discarding previous input
QUIT	stop current processing; begin again at start of program
STOP	end the job
MONITOR	return to the interactive monitor

The following presents a simple REAP scenario for retrieval of housing data for the nine counties of the San Francisco Bay Area. For convenience of understanding, the user input from the teletype has been presented in boldface.

```
WELCOME TO REAP
ENTER FILE NAME
HELP
  LIST OF AVAILABLE FILES

LBL-NAME   QUICK QWERY NAME     DESCRIPTION

CCDBMRG    CCDB.MERGE          CITY-COUNTY DATA BOOK (1952-1972)
CNAGR49    49.AGR.SUB          1949 U. S. CENSUS OF AGRICULTURE
CNAGR59    59.AGR.SUB          1959 U. S. CENSUS OF AGRICULTURE
CNAGR64    64.AGR.SUB          1964 U. S. CENSUS OF AGRICULTURE
CNAGR69    69.AG.CENCOM        1969+1964 U. S. CENSUS OF AGRICULTURE
CTY60PC    COUNTY.60.PC        1960 CENSUS OF POPULATION
OBERS      OBERS               SERIES C OBERS PROJECTIONS 1970
OBERSE     OBERS-E             SERIES E OBERS PROJECTIONS 1972
INCOME     INCOME              BEA INCOME TIME SERIES 1929-1969
ENERGY     ENERGY              COUNTY ENERGY DATA BASE
CNS4CT     CENSUS.4CT          FOURTH COUNT CENSUS (1970)
ENTER FILE NAME
CCDBMRG
 FILE=CCDBMRG-CCDB.MERGE       CITY-COUNTY DATA BOOK (1952-1972)
DEFINE STUDY AREA
NEW OR OLD?
NEW
ENTER STUDY AREA NAME
BAY AREA
ENTER 5 DIGIT FIPS STATE-COUNTY CODES
06001,06013,06041,06055,06075,06081,06085,06095,06097;
DISPLAY?
STATE.ABBR,COUNTY.NAME,52.CCDB048,52.CCDB049,62.CCDB051,72.CCDB077;
RETRIEVING DATA FROM MSS, PLEASE BE PATIENT
.
.
.
 STATE    COUNTY                1940       1950       1960       1970
 ABBR      NAME               DWELLING   DWELLING   HOUSING    HOUSING
                                UNITS      UNITS      UNITS      UNITS
CALIF   ALAMEDA                173031     247145     310312     379723
CALIF   CONTRA COSTA            31297      90641     124279     178329
CALIF   MARIN                   16472      28598      49581      70570
CALIF   NAPA                     8752      13746      21176      26791
CALIF   SAN FRANCISCO          222176     265726     310559     310383
CALIF   SAN MATEO               37230      76551     141770     190147
CALIF   SANTA CLARA             56406      91670     199922     336192
CALIF   SOLANO                  15312      33007      41894      53460
CALIF   SONOMA                  26831      39966      59784      77214

*TOTALS*                      ---------------------------------------

                              587507     887050    1259277    1622809

MORE QUERIES ON SAME STUDY AREA?
MONITOR
ENTERING MONITOR, PLEASE BE PATIENT
```

5. TECHNICAL DETAILS

The REAP data base is stored on a unique mass storage hardware device, the IBM-1360 photodigital mass storage system (known as the chipstore) with an online capacity of 50 billion bytes [7,8,9]. The only existing chipstore systems are at LBL, the Lawrence Livermore Laboratory and the Los Alamos Scientific Laboratory. This device is random access addressible to the 200 character segment.

The basic scheme was to use a locally written data base management system [6] called STOFI (for storage file) to manage a directory of addresses for the chipstore. The STOFI system was built to efficiently manage large disk-based data bases. STOFI typically manages 20 million byte files but is working effectively on a 100 million byte file. Ten basic subroutine calls were needed to build the large scale data access system around STOFI.

The system operates in three parts as illustrated in Figure 1. Initially the file or files are stored as data sets on chipstore. Once stored, they are read sequentially and a directory to the records is built based on a preprogrammed hierarchy set up by the user. The data dictionary, and the interactive help commands are also built into the directory file at this time. The third portion of the system is the interactive retrieval program, REAP.

The STOFI system allows logical data structuring in HISAM (Hierarchical Index-Sequential Access Method) similar to IBM's IMS data management system. The structure is illustrated in Figure 2.

The salient features of the system are

1. Data bases are stored in sequential order. The logical file structure is maintained independent of the actual data.

2. The logical structure is a hierarchy of keys taken from the data element values within each record. Keys may be specified either by name (taken from the data dictionary for the file), by position (location within the record) or by user supplied values.

3. Data dictionaries are stored on the index file and used to display individual fields in a record.

4. The interactive code is data driven such that, help commands and promptings will depend on the structure of the user's file directory.

5. Help aids, user lists and job control streams are stored on the user's directory file.

6. The data is archival in nature, i.e., the system is a retrieval not an update system.

For further details the reader is referred to reference [3].

6. BATCH RETRIEVAL

For some applications it is desirable to retrieve more information from a file than is practical to view at an interactive terminal, and in some instances one might wish to build a new data base at the record level to be run on a computer at an installation other than LBL. A program BREAP (Batch REAP retrieval) has been implemented whereby the user may extract from a given file only those records within his realm of interest. The BREAP language is compact and free format. For example the lines -

```
FILE = CNAGR69
CODES = 06001-06999
END
```

will retrieve (from chipstore to disk) all county records within the state of California for the 1969 U.S. Census of Agriculture.

Records extracted using BREAP may then be used as input to other programs and on other computers. The program has been used for data base file transfer across different computers at the LBL computer center.

7. DATA EXTRACTION AND NEW FILE CREATION

One of the more important tasks within the socio-economic-environmental-demographic field is the extraction of data for modeling, graphical displays, mapping, and statistical analysis. With this view in mind, a batch program FILMERG was designed which will retrieve data at the record and data element levels from up to ten different files thus allowing the user to form a new data base consisting only of those records and data elements which meet his specific needs. An example of input to FILMERG is -

```
FILE = CCDBMRG, DE = STATE.ABBR, COUNTY.NAME, 72.CCDB003,
72.CCDB058
FILE = ENERGY, DE = C4, C147, C67, C93
FILE = CNAGR69, DE = 69.AGT01008, 69.AGT01010;
CAT = 3*1
CODE = 3*06001, 3*06013, 3*06041, 3*06055, 3*06075, 3*06081, 3*06085,
3*06095, 3*06097
```

The above example opens the data base and retrieves data for the nine Bay Area counties from the City County Data Book, the County Energy Data Base, and the U.S. Census of Agriculture (1969 + 1964). The data elements of the newly created (merged) data base consist of the state abbreviation, the county name, the total population in each county in 1970, the median family income, the total county personal

income, the population density, the barrels of gasoline used in the county, the total electricity consumption, the approximate acres of land area in 1969, and the percent of land in farms.

Applications of FILMERG

The program has been used to generate a national county-level data base merging 600 elements from 3 files (City County Data Book, 1969 Census of Agriculture, and BEA Income). The new data base is being used at the Department of Interior computerized data center in Denver, Colorado.

In addition, Figure 3 is a map generated from FILMERG-extracted data plotted by the CARTE thematic mapping program. This data and mapping capability is being used by LBL's Energy and Environment division as part of an impact study on the current California drought funded by the California Energy Resources, Conservation and Development Commission.

8. REPORT GENERATION

Since report producing is an integral part of data collection, a program YARG (Yet Another Report Generator) was designed to operate as a completely compatible front-end processor to an in-house LBL text formatting program. Taking advantage of the capabilities of the text processor, YARG provides

- Automatic placement of $ (for monitary values) and commas (for demographic data)
- Automatic table of contents and indexing
- Mixed textual and tabular reporting
- Automatic right and left justification

Currently YARG is a report generator only; it offers only data retrieval from existing REAP-style files and may be run in an 'online batch' mode in conjunction with BREAP. Sitting at an interactive terminal, the user may define the georgaphic areas of interest, retrieve the data records through batch mode, write a report which requires selected data elements within the records, and produce a final report with the desired data inserted.

Figure 4 illustrates input to create the YARG report for California Agriculture, a portion of which is displayed in Figure 5.

9. PLANS FOR FUTURE DEVELOPMENT

Within the next year, the following files should be added to the data base -

.1972 Census of Mineral Industries
.1974 Census of Agriculture
.1960-1970 County migration by Age, Sex, and Race
.County Cancer Mortality Data by Sex and Race

Plans to extend the methodology to retrieve large data bases stored on 6250 cpi tape will be implemented. In addition, some consideration is being given to a minicomputer version of REAP.

10. ACCESS / ACKNOWLEDGMENTS

REAP is available via dial-up terminal to U. S. Federal government agencies or other users on Federally funded projects. In addition, REAP is available through the ARPANET connection at the Lawrence Berkeley Laboratory computing facility. For further details, contact the authors. REAP development was funded primarily by the U. S. Army Corps of Engineers (Planning Division), with additional funding from the U. S. Department of Labor (Employment and Training Administration) and the U. S. Bureau of Reclamation, under interagency agreements with the Energy Research and Development Administration.

11. REFERENCES

[1] Austin, Donald, Suzanne Kranz, and Carl Quong, An Overview of the LBL Socio-Economic-Environmental-Demographic Information System, Report LBL-3699, Berkeley, March, 1975.

[2] Wood, Peter, Interactive Thematic Mapping, A Report, Presented at URISA 76, Atlanta, October, 1976, Report LBL - 5362, Berkeley, October, 1976.

[3] Gey, F. C., and Marilyn Mantei, Keyword Access to a Mass Storage Device at the Record Level, Proceedings of the First International Conference on Very Large Data Bases, Boston, September 1975.

[4] Kreps, Peter, and Virginia Sventek, SEEDIS Monitor, Report LBL-6440, Lawrence Berkeley Laboratory, Berkeley, June 1977.

[5] Calef, Charles, and Frank Drysdale, The Energetics of the United States, An Atlas, Report BNL - 50501, Brookhaven National Laboratory, Upton, New York, July 1976.

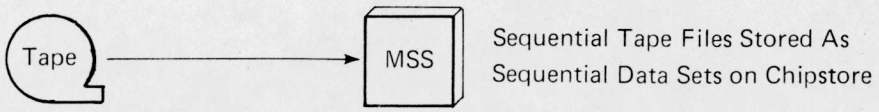
[6] Mantei, M. and Sarna, J., An introductory manual to STOFI, report LBL-2649, LBL, Berkeley, CA (1974).

[7] Penny, S.J., Fink, R. and Alston-Garnjost, M., Design of a very large storage system, Proc. of AFIPS FJCC, 37 (1970) 45-51.

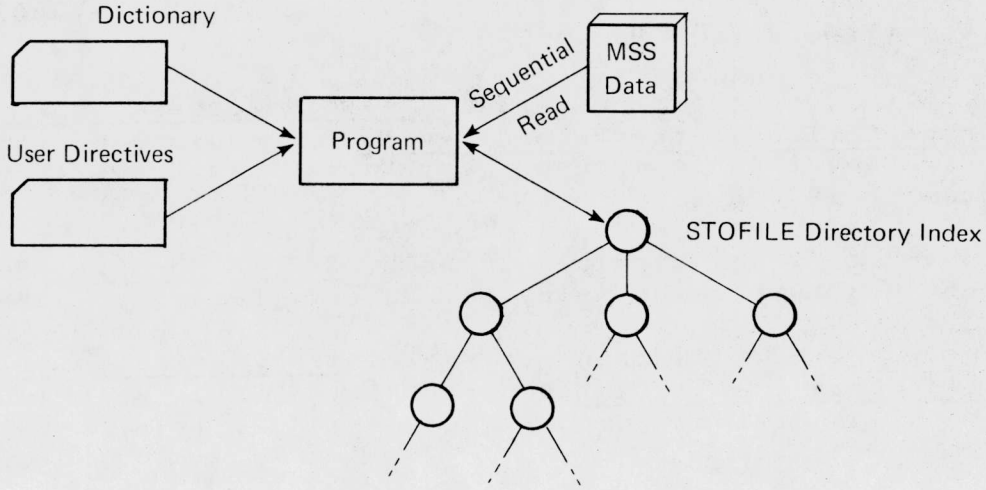
[8] Kuehler, J.D. and Kerby, H.R., A photo-digital mass storage system, Proc. of AFIPS FJCC, 29 (1966) 735-742.

[9] Furman, R.M., IBM 1360 photo-digital mass storage system, IBM technical report, Tr 02.427 (May 15,1968).

Part I. Data Storage



Part II. Directory Creation



Part III. Data Retrieval

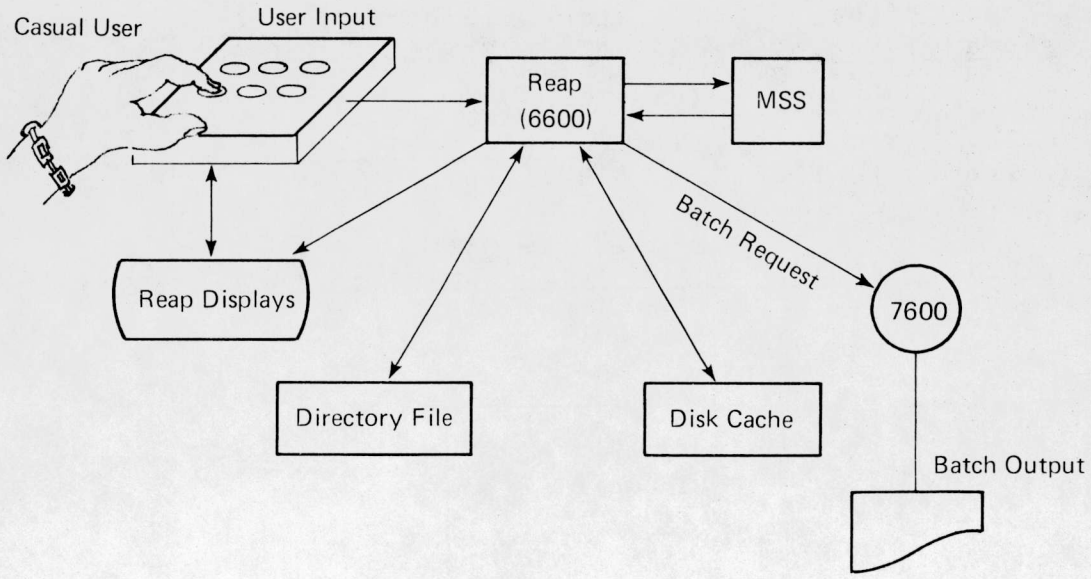


Figure 1 Tripartite System Operation

XBL 758-3719

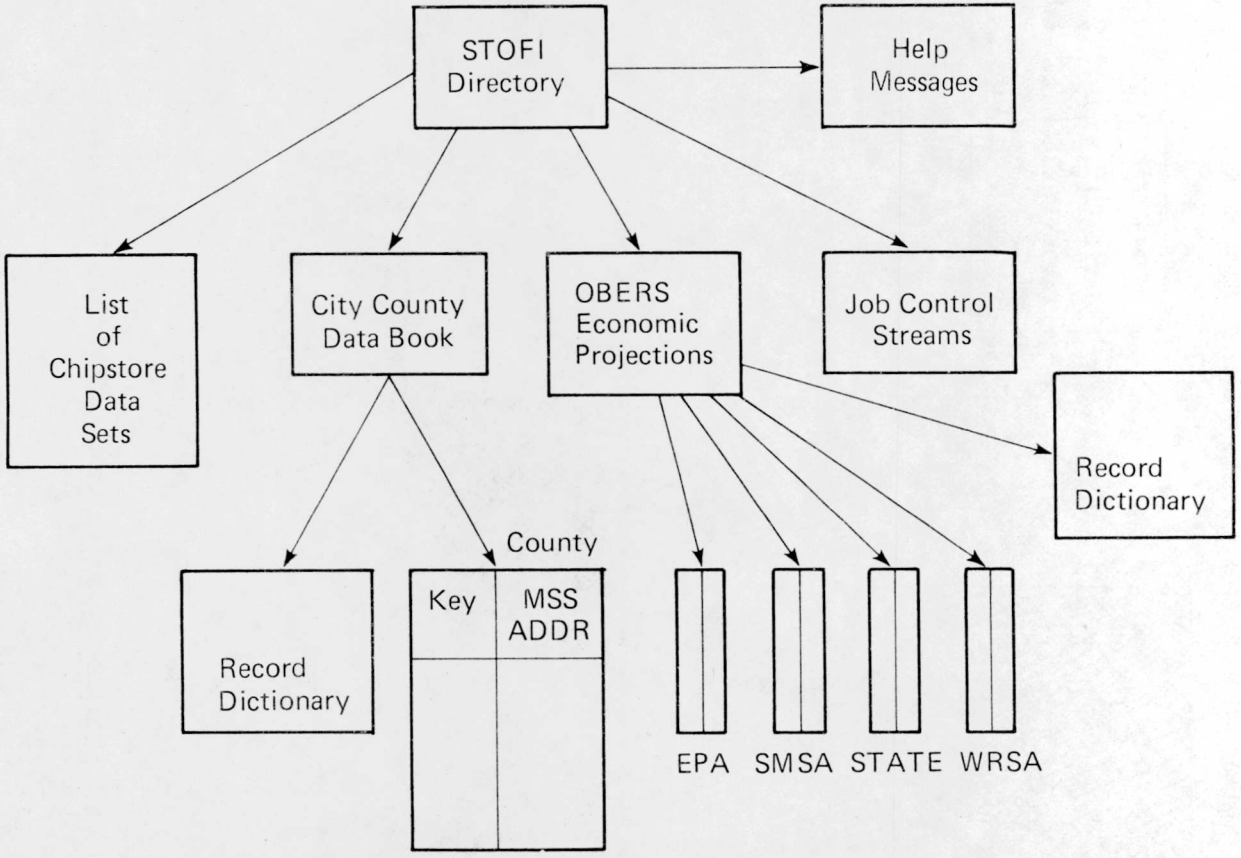


Figure 2 Sample STOFI Directory file for US Economic Data
(City-County Data Book and OBERS Projections)

XBL 758-3722

PAGE 1,COLOR 1

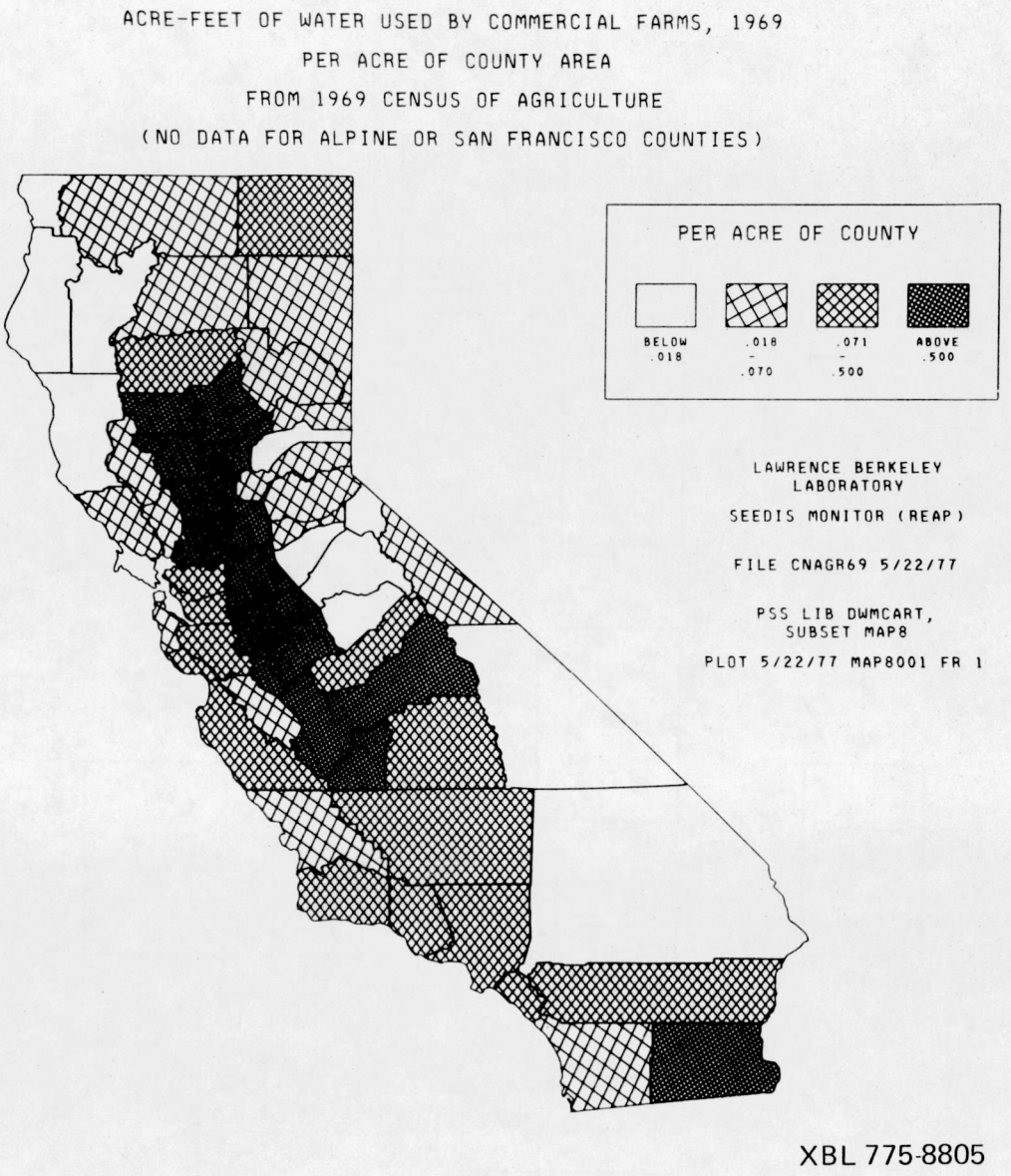


Figure 3: FILMERG output plotted by CARTE

```
M5,75,60,72
ZL2
ZF
#  *
6
W SUMMARY !A!GRICULTURE !R!EPORT
2
W !C!OUNTIES WITHIN !C!ALIFORNIA
2
W 1969 AND 1964 !U!. S. CENSUS OF !A!GRICULTURE
E
T SUMMARY AGRICULTURE CENSUS REPORT
  **
H *COUNTY.NAME* COUNTY, !C!ALIFORNIA
P IN 1969 WITHIN THIS COUNTY, THERE WERE *69.AGT01001* FARMS.  THE TOTAL
 LAND AREA IN FARMS WAS *,69.AGT01003* ACRES, WHICH COMPRISED
 *69.AGT01010* PERCENT OF THE TOTAL LAND AREA IN THE COUNTY.
 THE AVERAGE FARM SIZE WAS *69.AGT01005* ACRES, AND THE AVERAGE VALUE
 OF LAND AND BUILDINGS PER FARM WAS *$,69.AGT01013*.
P THE FOLLOWING SUMMARIZES CROP PRODUCTION IN 1969 AND 1964 FOR THE
 COUNTY.
2
F_L40,C30
K_TABLE 1 - !C!ROP !P!RODUCTION
K_AMOUNT HARVESTED
F_L40,R15,R15
K
U !C!ROP _1964 _1969
2
K
N CORN USED FOR GRAIN (BU)_*,64.AGT08007*_*,69.AGT08007*
K
N SORGHUMS FOR GRAIN OR SEED (BU) _*,64.AGT08021* _*,69.AGT08021*
K
N WHEAT FOR GRAINS (BU) _*,64.AGT08033*_*,69.AGT08033*
K
N SOYBEANS USED FOR BEANS _*,64.AGT08043* _*,69.AGT08043*
K
N HAY (TONS) _*,64.AGT08006*_*,69.AGT08006*
K
N COTTON (BALES) _*,64.AGT08010*_*,69.AGT08010*
K
N PEANUTS FOR NUTS (POUNDS) _*,64.AGT08015*_*,69.AGT08015*
K
N TOBACCO (POUNDS) _*,64.AGT08022*_*,69.AGT08022*
E
Z
```

FIGURE 4 - INPUT TO YARG REPORT

Summary Agriculture Report

Counties within California

1969 and 1964 U. S. Census of Agriculture

ALAMEDA COUNTY, CALIFORNIA

In 1969 within this county, there were 762 farms. The total land area in farms was 291,055 acres, which comprised 62.0 percent of the total land area in the county. The average farm size was 381.8 acres, and the average value of land and buildings per farm was $297,820.

The following summarizes crop production in 1969 and 1964 for the county.

Table 1 - Crop Production
amount harvested

Crop	1964	1969
Corn used for grain (bu)	27,116	520
Sorghums for grain or seed (bu)	21,532	12,677
Wheat for grains (bu)	13,270	2,773
Soybeans used for beans	0	0
Hay (tons)	28,166	27,799
Cotton (bales)	0	0
Peanuts for nuts (pounds)	0	0
Tobacco (pounds)	0	5

AMADOR COUNTY, CALIFORNIA

In 1969 within this county, there were 206 farms. The total land area in farms was 257,821 acres, which comprised 69.0 percent of the total land area in the county. The average farm size was 1251.5 acres, and the average value of land and buildings per farm was $238,672.

The following summarizes crop production in 1969 and 1964 for the county.

Table 1 - Crop Production
amount harvested

Crop	1964	1969
Corn used for grain (bu)	0	12,031
Sorghums for grain or seed (bu)	3,214	16,949
Wheat for grains (bu)	557	1,431
Soybeans used for beans	0	0
Hay (tons)	4,128	7,249
Cotton (bales)	0	0
Peanuts for nuts (pounds)	0	0
Tobacco (pounds)	0	0

Figure 5: YARG report example

BUTTE COUNTY, CALIFORNIA

In 1969 within this county, there were 1844 farms. The total land area in farms was 551,254 acres, which comprised 52.3 percent of the total land area in the county. The average farm size was 298.8 acres, and the average value of land and buildings per farm was $168,334.

The following summarizes crop production in 1969 and 1964 for the county.

Table 1 - Crop Production
amount harvested

Crop	1964	1969
Corn used for grain (bu)	156,243	127,316
Sorghums for grain or seed (bu)	853,170	542,265
Wheat for grains (bu)	126,298	229,507
Soybeans used for beans	0	0
Hay (tons)	46,338	48,100
Cotton (bales)	0	0
Peanuts for nuts (pounds)	0	51,292
Tobacco (pounds)	0	0

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720