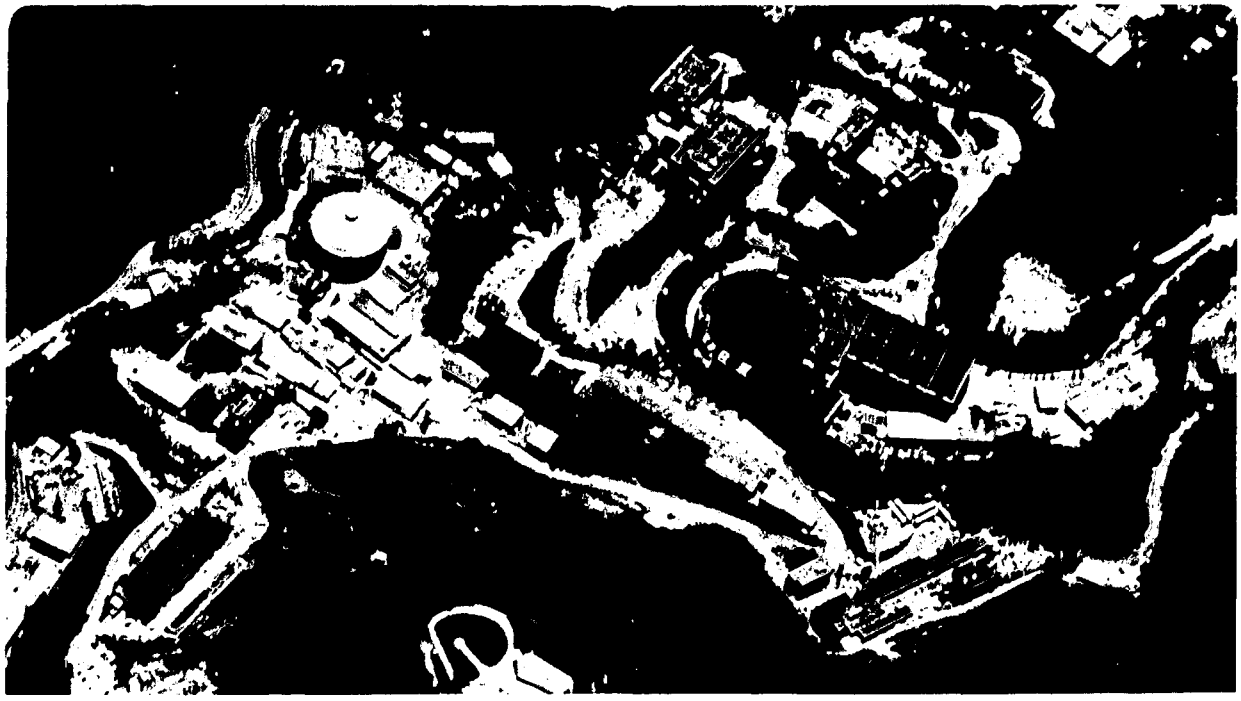
UC-13
LBID-379

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Physics, Computer Science & Mathematics Division

For Reference

Not to be taken from this room



DISCLAIMER

lestz — for TID

THE SEEDIS PROJECT: A SUMMARY OVERVIEW

John L. McCarthy
Aaron Marcus
William H. Benson
Deane W. Merrill
Fredric C. Gey
Carl Quong

DRAFT

Computer Science and Mathematics Department
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

April 1981
LBID 379

This work was supported by the U.S. Department of Energy under Contract Number W-7405-ENG-48.

SEEDIS is a research and development project on Social, Economic, Environmental, and Demographic Information Systems at the Lawrence Berkeley Laboratory (LBL), supported by the Department of Energy, Department of Labor, and others. The SEEDIS project includes:

- an LBL Computer Science and Mathematics Department research program on distributed, interactive information systems
- an integrated, interactive, distributed, testbed information system running on a network of VAX computers, which is used for selected applications as well as research and development
- a set of information management and analysis tools for research applications in fields such as energy management, water resource evaluation, manpower planning, and epidemiology.
- a major collection of databases for various geographic levels and time periods drawn from the United States Census Bureau, Department of Energy, Department of Labor, Environmental Protection Agency, National Center for Health Statistics, Bureau of Economic Analysis, and other sources.

PURPOSE

Policy formulation, implementation, and management depend upon accurate, timely information. Policy makers, managers, analysts, and technicians need tools to locate, retrieve, combine, analyze and display information from a variety of sources. For most decision-making purposes time and resources usually do not permit collecting new data, but there is a wealth of publicly available government and private data that often could meet such needs if it were quickly and easily accessible.

Unfortunately, despite the fact that computers and machine-readable data have made it potentially easier to locate and analyze information, it is still quite difficult to find and use specific data items of interest. Combining information from different sources which may reside in different physical locations adds further complications. Even within major data archives such as the United States Census Bureau, National Center for Health Statistics, Environmental Protection Agency, and the Inter-University Consortium for Political and Social Research, reliance on printed indexes limits access to machine-readable data. Furthermore, it is difficult if not impossible to combine multiple datasets because of differing code conventions, data structures and units of analysis.

The SEEDIS Project addresses these basic problems through research, design, and development of distributed information system components. In particular, VAX SEEDIS provides a unified framework for data management, information retrieval, statistical analysis, and graphical display. Using SEEDIS, non-programmer users can efficiently access and manipulate very large, diverse, and distributed statistical databases. In some of these respects, SEEDIS resembles systems such as Statistics Canada's RAPID DBMS [STAT 77], the Decision Information Display System (DIDS), developed by NASA and the Department of Commerce [DALT 79, DECI 81], and UPGRADE, developed by the President's Council on Environmental Quality) [COUN 80].

SEEDIS project staff work with people in selected applications in order to

- implement and evaluate information system components
- acquire and develop new databases
- test the viability of new concepts and tools in a "real world" large database environment
- get feedback from knowledgeable subject area specialists about how information systems tools can be improved

PROJECT HISTORY

Although VAX SEEDIS is only two years old, the motivation and experience which led to its development spans nearly ten years. In 1972, the Department of Labor asked the Lawrence Berkeley Laboratory (LBL) to apply its expertise with very large databases from accelerator experiments to development of storage, retrieval, and report generating software for 1970 United States Census data. This effort subsequently led to development of software for interactive access to the growing collection of databases, tools for mapping and graphic display [GEY 75, WOOD 78], plus an interchange file system and command language "monitor" to link the various evolving subsystems on CDC 6000 series machines [AUST 75]. Work on the integrated VAX version of SEEDIS began in 1979, and efforts are currently underway to add major enhancements necessary for incorporation of 1980 census data [COMP 81, GEY 81, MARC 81, MERR 80].

FEATURES

Unlike many other statistical information systems, SEEDIS provides a testbed for different functional components as they become available, including software developed at LBL and elsewhere. Its underlying file interchange format and command language interpreter are designed to provide a "software bus" for interchange of data and data descriptions among a variety of storage and access methods, search and retrieval tools, display and analysis facilities, and user interface environments -- so that users need not be concerned with the detailed structure or operating requirements of indvidual system components. Exhibit 1 presents the logical structure of SEEDIS, with its underlying file interchange format, unified user interface, and various functional modules.

Major features of the current version of SEEDIS, as ellaborated in the subsections below, include the following:

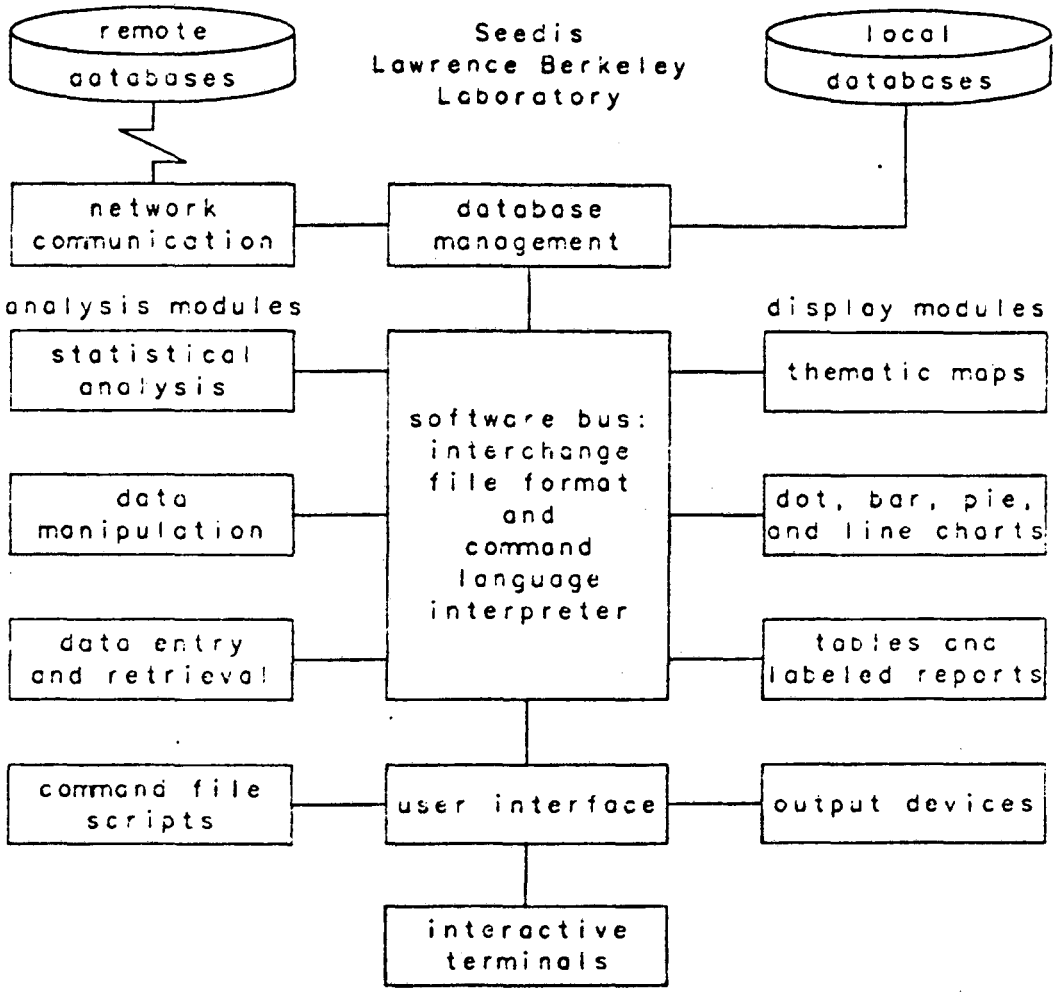
- efficient handling of very large numeric databases
- distributed operation over a network of VAX computers
- a "user friendly" human-computer interface
- flexible selection and manipulation of data
- interactive color chart and map making facilities

Special Facilities for Large Numeric Databases

Because many SEEDIS databases are quite large, considerable attention has been given to methods for efficient compression, storage, and retrieval of numeric data. SEEDIS currently uses a computer independent binary compression technique based on run length encoding of zeros and missing data, which has required only twenty percent of the original storage space for census data. [????] New techniques developed by CSAM staff promise even greater storage and retrieval efficiency. [EGGE 81]

Excluding duplicate copies, SEEDIS databases currently occupy approximately 25 billion bytes of storage, primarily in compressed form on some 250 high density (6250 bpi) magnetic tapes. With the addition of data from the 1980 census and other sources, the collection will probably double in the next two years. In printed form, the data would occupy over 25 million pages. Long-range SEEDIS development plans call for fast, interactive access to the complete set of databases, using video disk mass storage and distributed data management techniques.

Exhibit 1: A Schematic View of the Major Functional Components of SEEDIS:



At present, a prototype subset of 120 databases (300 megabytes of the most frequently accessed data) is stored on disk for immediate access from the VAX version of SEEDIS. This prototype set of databases currently contains a total of 22,000 different data items (a total of over 40 million data values). It includes 6800 different data items for each of the roughly 3000 counties in the United States, and 1600 different data items for each state.

Other data, such as the large fourth, fifth, and sixth count 1970 census files for small geographic areas (e.g., enumeration districts, block groups, tracts, minor civil divisions), was formerly stored on a special photodigital storage device (the IBM 1360 "chipstore") and accessed via special purpose programs on a CDC 6600 mainframe computer. It is currently stored in computer independent binary format on high density tapes, which can be read on any type of hardware using special SEEDIS routines.

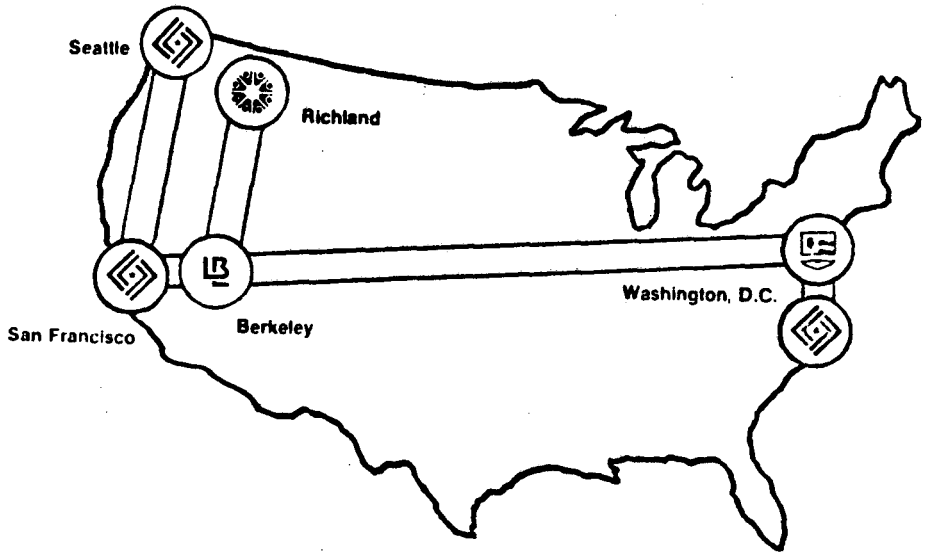
Within the next year, VAX SEEDIS users will be able to access all of the data directly via a network link to the LBL Computer Center's tape robot General Storage System (GSS). This facility will provide automatic access (with an average of extraction time of 2 to 30 minutes) for moderate (up to 10 megabyte) subsets of data until fast-access, low-cost mass storage devices become available

Distributed Network Facilities

One of the goals of the SEEDIS project is to give users shared access to a wide variety of databases stored in different physical locations, while permitting each local facility to maintain control over its own databases. Eventually, SEEDIS will provide distributed data management, retrieval and analysis capability over networks of hetereogeneous computer systems.

SEEDIS presently operates in a network of seven DEC VAX minicomputers. This distributed computer network (DCN), is pictured below in Exhibit 2. The primary VAX SEEDIS system and data are stored on disk drives connected to two VAXes in Berkeley, California. Other nodes each have 25 megabytes of heavily used SEEDIS program modules and data description files physically resident on local disk. Network facilities provided by DEC-net [DIGI 79] enable users at any node to access databases anywhere on the network. Except for response time, SEEDIS behaves as if all the data were stored locally.

Exhibit 2: The Distributed Computer Network, 1981



Department of Energy Research Laboratories
- Berkeley: Lawrence Berkeley Laboratory
- Richland: BATTELLE Pacific Northwest Laboratories
- Washington, D.C.: George Washington University

Department of Labor, Employment & Training Administration
- Seattle: Federal Region X Office
- San Francisco: Federal Region IX Office
- Washington, D.C.: National Office

User Interface

During the past year, SEEDIS staff have put substantial efforts into designing consistent vocabulary, layout, and sequencing of system dialogue and display in order to improve the effectiveness of SEEDIS from a human factors point of view [MARC 81]. A sample dialogue appears in appendix A below. Current features of the SEEDIS user interface include:

- interactive operation with menu prompting
- online help, explanation of commands, data catalogs, and status information
- searching aids for data selection, including online browsing of data dictionaries
- standard socio-economic report formats (profiles) for user-defined areas
- optional batch task submission
- logging of user dialog, which can be used for locating problems and creating batch procedures

Data Selection and Manipulation

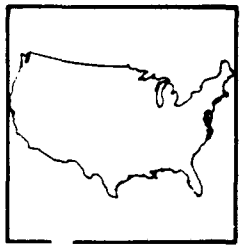
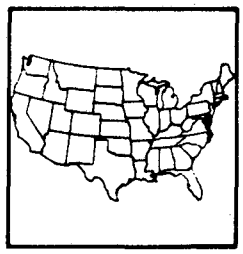
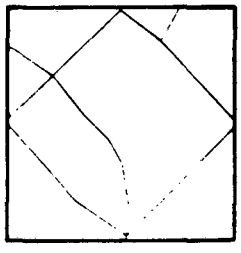
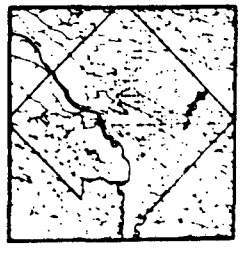
SEEDIS includes a number of powerful and unique features for selecting level of analysis (e.g., census tracts, counties, or one of fifty other levels -- see section ?? below); scope of analysis -- specific units or entities within the selected level (e.g., all counties in Federal Region IX, selected tracts in New York City, etc.); and particular data items of interest (e.g., number of hispanic families with annual income less than $3000, death rate from leukemia, total amount of suspended particulates for 1976, etc.)

Exhibit 3 illustrates the concepts of geographic level and scope for some of the major geographic areas currently implemented in SEEDIS. For example, the lower right corner pictures specification of census tract level with a scope of all tracts in the District of Columbia. The example SEEDIS-user dialogue in appendix A shows the actual commands used to select state level data for several states.

Other SEEDIS facilities enable users to manipulate data and entities to which the data pertains in a variety of ways, including the following.

- subselect data and particular geographic entities on the basis of data item values (e.g., census tracts in which the proportion of housing units with oil heat exceeds fifty percent)
- automatically aggregate, disaggregate, interpolate and integrate data from different geographic levels into a single analysis file
- create new data items, sets of entities, etc. using logical and arithmetic functions
- produce self-documenting intermediate data files for use in subsequent SEEDIS sessions or software external to SEEDIS

Exhibit 3: Examples of Geographic Level and Scope in SEEDIS

SCOPE	LEVEL			
	NATION	STATE	COUNTY	TRACT
WORLD				
U.S.				
EASTERN U.S.				
D.C. AREA				
D.C.				

SOCIO-ECONOMIC ENVIRONMENTAL DEMOGRAPHIC INFORMATION SYSTEM
LAWRENCE BERKELEY LABORATORY

Interactive Graphics

SEEDIS incorporates a number of capabilities from the forefront of computer graphics research as well as standard facilities. SEEDIS mapping facilities have been used to produce major cartographic publications such as the Urban Atlas, which was a joint effort of LBL and the United States Census Bureau [ ]. Current graphics features include the following:

- Production of graphic displays on a variety of standard monochromatic and color devices
- Custom labeled tables, bar charts, pie charts, line graphs, scatterplots, and other graphic output [EADE 81] (see Appendix B for examples)
- Polygon (choropleth) and symbol mapping for predefined geographic entities [YEN 79] (see Appendix C for examples)
- Special color maps and charts, including bivariate displays [TRUM 80] and "fuzzy graphics" [BENS 81] (~~examples of which appear in~~ Appendix D)

DATABASES

Since the early 1970's, the SEEDIS Project has acquired and developed several hundred different databases containing over 170 thousand different data items (roughly three billion individual data values). In addition to files from the 1970 United States Census, which account for about half of the current data inventory, SEEDIS holdings also include energy, health, demographic, environmental, and socio-economic databases obtained in connection with a variety of applications projects [BURK 79B]. Exhibit 4 outlines the current inventory of SEEDIS databases, along with summary information about database contents, sizes, geographic coverage, and accessibility.

Each data item (i.e., a variable such as the number of unemployed persons, or a table such as unemployment by age, race, and sex) is available for one or more distinct geographic levels (e.g. county, state, etc.). Each SEEDIS database contains information for a set of comparable geographic entities defined at the same level (e.g., counties as defined in the 1980 census, or Standard Metropolitan Statistical Areas ~~(SMSA's)~~ as defined in 1979).

Over fifty different geographic levels, from nations down to census tracts, are presently defined in SEEDIS. New levels are defined as needed to accommodate new files. Multiple databases may exist for any given geographic level, and users can combine data from different databases and levels for purposes of display and analysis

too much space. 1 tabs for cols drop font

Exhibit 4 Summary of Major SEEDIS Databases

Database Title or Description	Year(s) Covered (1)	Data Cells /Variables (2)	Summary Level(s) (3)	M. Data Values (4)
MAJOR VAX SEEDIS DISK FILES				
Social and Demographic Characteristics				
County Data Book	47-77	1022	sc	3.3
Census 4th Count, by race	70	5890	sm	1.7
Survey of Income & Educ Tabulations	76	4103	s	0.2
Population by age, race, sex	70-77	608	nsmco	4.4
Economy and Employment				
BLS Labor Force and Unemployment	74-79	72	sco	0.1
BEA Economic Projections	69-2030	560	nrs	0.6
Census of Agriculture	74	1200	sc	3.8
Employment by Type and Industry	71-76	248	sc	0.9
Energy and Environment				
1974-76 Air Quality	74-76	257	o	0.8
Air Quality Monitor Station Directory	74-76	59	o	0.4
1960-1995 Electric Generating Capacity	60-95	18	o	0.1
Epidemiology and Health				
Age Specific Mortality by Race and Sex	68-72	198	nrsc	0.7
Life Expectation	68-72	44	nrsc	0.2
Leukemia Mortality	69-71	36	nrsc	0.2
Cancer Mortality	50-69	424	co	1.3
Cancer Incidence by Site and Histology	69-71	352	t	1.6
Age Adjusted Mortality	68-72	652	c	2.0
Area Resource File	77	889	c	1.3
Total VAX SEEDIS Disk Files		22,738		31.9
MAJOR SEEDIS GENERAL STORAGE SYSTEM TAPE FILES				
1970 Census				
First Count	70	400	smcdptbo	100
Second Count	70	3,500	smcdpto	100
Fourth Count	70	6,000	smcdpto	450
Fifth Count	70	800	smcdptbo	300
Sixth Count	70	150,000	smcdptbo	150
Public Use Samples	70	400	ih	200
County Business Patterns	64-73	10,000	nsc	300
Current Population Surveys	70-79	300	ih	200
Employment by Industry and Occupation	70	800,000	nh	100
Survey of Income and Education	76	490	ih	100
Other Miscellaneous Files				400
Total SEEDIS GSS Tape Files		972,000+		2,400

(1) Some files are annual series while others cover only selected years for certain variables. For information on specific years in each series see online data dictionaries or [BURK 79B]

(2) Number of distinct data cells for aggregate data; number of variables for household and individual level data. Number available for a single year in cases where time series are available.

(3) Major geographic levels for which data are available, coded as follows:

n nation	p places
r interstate regions	t tracts
s states	b block groups/ennumeration dists
m standard metro stat areas	h households
c counties	i individuals
d minor civil divisions	o other

(4) Millions of individual data values (i.e., number of variables or data cells times number of summary unit records times years).

GEOGRAPHIC AREA AND MAP FILES

In order to facilitate combination of data from different levels of analysis, special geographic files in SEEDIS define each geographic unit in terms of the larger entities of which it is a part. For example, every county (1970 census definition) is identified as belonging to a particular EPA Air Quality Control Region, a particular Bureau of Economic Analysis Area, etc. Where necessary, counties are divided into smaller undivided units whose assignment to larger areas is uniquely defined. Exhibit 5 summarizes the current list of SEEDIS geographic levels and the number of individual entities (areas) in each. SEEDIS provides facilities for users to browse such lists online for easy reference.

For mapping purposes, SEEDIS also includes a set of cartographic base files -- one for each geographic level [BURK 79A]. Each geographic unit within a given level is associated with a series of latitude-longitude coordinate pairs, which define a polygon representing its boundaries. Some polygons are aggregates of county polygons; others, corresponding to subcounty areas, were carved out of county polygons. Point locations such as oil pipeline terminus points or air quality monitoring stations are identified by a single latitude-longitude coordinate pair. All map files are archived in latitude-longitude coordinates, in order to permit overlaying of different geographic entities. Projection, for example to conic coordinates, is performed at run time as required for display purposes.

Exhibit 5 Major Geographic Levels Defined in SEEDIS as of 4/81

Geographic Level Description	Units in Level
INTERNATIONAL	
Nations (1980 FIPS definitions)	233
LARGE INTERSTATE	
Bechtel Energy Model Regions	14
Census Regions	9
Coal Supply Regions	12
Federal Regions	10
National Petroleum Council Oil and Gas Areas	12
Petroleum Allocation Districts	7
1979 Standard Consolidated Statistical Areas	13
Water Research Council Regions	22
SMALL INTERSTATE	
Bureau of Economic Analysis Areas	
1969	173
1977	183
Bureau of Labor Statistics Labor Market Areas	437
1970 Census Public Use Sample County Groups	408
EPA Air Quality Control Regions	247
Standard Metropolitan Statistical Areas	
1971	247
1973	267
1975	276
1979	288
New England County Metropolitan Areas	276
Water Resources Subareas	222
STATE AND SUBSTATE	
States and Territories	55
1970 State Economic Areas	510
Single state portions of various interstate regions	4397
1980 Bureau of Labor Statistics Prime Sponsors	469
COUNTY AND SUBCOUNTY	
Counties	
1970 Census	3255
1980 Census	3257
Johns Hopkins Mortality Survey Program	3075
National Center for Health Statistics	3082
National Cancer Institute	3061
Places	
1980 Bureau of Labor Statistics	1565
Environmental Protection Agency	9745
1970 Census, Population over 1000	11970
Minor Civil Divisions	
1970 Census	35198
Tracts	
1970 Census	34648
POINT LOCATION	
1974-76 Air Quality Monitoring Stations	6625

6. Availability

SEEDIS is currently being used by the United States Department of Labor, Department of Energy, Environmental Protection Agency and Army Corps of Engineers. Other organizations or individuals interested in using the system have several alternatives, as follows:

- The National Technical Information Service prepares standard reports based on 1970 census data for user-designated census areas or aggregations thereof. For information, write or call:

  Marvin Wilson, NTIS
  5285 Port Royal Road
  Springfield, VA 22161
  (703) 487-4805, (FTS) 737-4805

- The State Data Program/Survey Research Center on the University of California's Berkeley campus provides standard reports similar to those of NTIS as well as more specialized data extraction services at cost. For information, write or call:

  Ilona Einowski, Data Librarian
  SDP/SRC 2538 Channing Way
  University of California
  Berkeley, CA 94720;
  (415) 642-6571

- In the future, VAX SEEDIS itself will be made available for distribution through the National Technical Information Service. Organizations interested in installing SEEDIS can contact either NTIS at the above address or

  Harvard Holmes, SEEDIS Project
  Computer Science and Applied Mathematics Department
  Lawrence Berkeley Laboratory
  Berkeley, CA 94720
  (415) 486-5181, (FTS) 451-5181

For further written information on SEEDIS, please see the references listed below.

REFERENCES

[AUST 75] Austin, D.M., Kranz, S.G., and Quong, C.: "An Overview of the LBL Socio-Economic Environmental Demographic Information System (SEEDIS)." Lawrence Berkeley Laboratory Report LBL-3699 (March, 1975).

[BENS 77] Benson, W.H.: "Interactive Analysis and Display of Tabular Data." 2 Computer Graphics 2 (Summer, 1977), 48-53

[BENS 81] Benson, W.H.: "An Application of Fuzzy Set Theory to Data Display" in R.R. Yager (ed.), Recent Developments in Fuzzy Set and Possibility Theory (Pergamon Press, forthcoming 1981)

[BURK 79A] Burkhart, B.R., ed.: Cartographic Base Files at Lawrence Berkeley Laboratory: 1978 Inventory; LBL-8707, January 1979.

[BURK 79B] Burkhart, B. and Merrill, D., eds.: Spatial Data on Energy, Environmental and Socioeconomic Themes at Lawrence Berkeley Laboratory: 1978 Inventory; LBL-8744, UC-13, April 1979. Included in Oak Ridge National Laboratory Report EIS-144, November 1978.

[COMP 81] Computer Science and Applied Mathematics Department, SEEDIS Release Notes (version 1.1) [in preparation]

Council on Environmental Quality, UPGRADE User's Manual, Washington, DC 1980

[DALT 79] Dalton, Billingsley, Quann, and Braker: "Interactive Color Map Displays of Domestic Information," 13 Computer Graphics 2 (August, 1979) 226-233

[DECI 81] Decision Information Display System Program Office, System Description and User's Guide, Washington, DC 1981 [in preparation]

[DIGI 79] Digital Equipment Corporation, DECnet-VAX User's Guide, Maynard Massachusetts 1979.

[EADE 81] Eades, Craig, CHART: A Graphic Display and Analysis System, A User's Guide, LBL-3015, 1981

[GEY 75] Gey, F. and Mantei, M. "Keyword Access to a Mass Storage Device at the Record Level," Proceedings of the First International Conference on Very Large Databases, Framingham, Massachusetts, September 1975 [GEY 75] Gey, F. and Williams, E., "The Reap Family of Computer Programs fro Retrieval of Socio-Economic-Environmental-Demographic Information", LBL-6417, June, 1977

[GEY 81] Gey, F., A Beginner's Guide to SEEDIS, LBL-11198, January 1981

[MARC 81] Marcus, A., [Style Manual], [in preparation]

[MERR 80] Merrill, Deane; CODATA Users' Manual; LBID-021, revised April 1980.

[STAT 77] Statistics Canada, RAPID DBMS, Ottawa, 1977.

[WOOD 78] Wood, Peter and William Benson, Computer Mapping Software at the Lawrence Berkeley Laboratory, LBL-7938, 1978.

[YEN 79A] Yen, Albert; VAX/CARTE User's Manual; 1979.

[YEN 79B] Yen, Albert and Wood, P.; "Moving Interactive Thematic Mapping from Mainframe to Mini: Some Design Possibilities and Development Experience", AUTO Carto IV Proceedings, vol. II (1979), 379-386

Appendix A Example of Interactive SEEDIS Dialog (with annotations in right margin)

Appendix B Examples of SEEDIS Interactive Chart Production Facilities

Appendix C Examples of SEEDIS Interactive Map Making Capabilities

Appendix D Examples of SEEDIS Special Interactive Graphics Facilities

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

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