The Development of Software to Facilitate Use of Archived Data Sets

by Josefina J. Card and Elizabeth A. McKean[1]
Sociometrics Corporation

A variety of forces have converged to promote research based on secondary analysis of existing social science data sets. A growing number of federal agencies have officially encouraged data sharing by requesting, often requiring, that their grantees place data sets collected with public funds in the public domain. At the same time, declines in university and federal research budgets have put expensive primary data collection out of the reach of many social scientists. Advances in microcomputer technology have allowed powerful data analyses to be performed quickly and economically. Data archive centers dedicated to the preparation of data sets for public use have also begun to emerge.

Several challenges await the data archivist working to provide users with clean, useable data for secondary analysis. The user must be provided with data of high quality, documented in clear and comprehensive fashion. While paper documentation retains its value, paperless (electronic) documentation is becoming increasingly important, in light of burgeoning use of the Internet and mass storage media such as the CD-ROM. Hand-in-hand with the growth of the national movement toward secondary data analysis comes the need for powerful yet user-friendly ways to search through the massive amounts of available data and documentation to retrieve studies or variables of interest. To reduce hard-disk storage burdens and statistical analysis time, such search and retrieval of variables would ideally be linked with data extract capabilities, so that analysis files containing only variables or cases of interest to the user can be created on demand.

This article documents the latest advancements of one data archive center, Sociometrics Corporation, in meeting these challenges. The Sociometrics Data Library currently houses five topically-focused data archives: the Data Archive on Adolescent Pregnancy and Pregnancy Prevention, the American Family Data Archive, the Data Archive of Social Research on Aging, the Maternal Drug Abuse Data Archive, and the AIDS/STD Data Archive. Together, these five data archives include over 200 data sets, which have been chosen for technical quality, scientific merit, substantive utility, relevance to social policy, demand for secondary data analysis, and disciplinary balance by a panel of experts in each archive's substantive field. Each data set in each topically-focused archive is made publicly available with a printed and bound user's guide and a standard set of machine-readable files—raw data, SPSS and SAS program statements that fully document the variables and values in the data file, an SPSS dictionary, and SPSS frequencies—with the explicit goal of providing the user with clear documentation and ready-to-use data files (see Card and McKean, 1993 for a discussion of standard file preparation).

This paper will focus on the recent development of three software features accompanying the data sets which simplify secondary analysis for social scientists. While the examples used will be drawn from the AIDS/STD Data Archive, the software described is generic to the data archives comprising the Sociometrics Data Library.

Software to Facilitate the Selection and Analysis of Variables

Search and Retrieval Software. The first feature, Search and Retrieval software, allows the user to examine the contents of an entire topically-focused data archive and retrieve variables using a variety of search strategies: (1) searches by full-text keyword, including variable names, variable labels (question descriptors), and value labels (response descriptors); (2) searches by substantive Topic and Type codes that have been assigned to each variable during the archiving process; and (3) searches by study name, author, or assigned data set number. Standard Boolean operators such as "and," "or," and "not" can be used to conduct any search.

Electronic Instrument-Variable Link. The second feature, an electronic link between study variables and graphic images of the data collection questionnaire, allows the analyst to select variables and view the instrument pages associated with the selected variables. Alternatively, the user may browse through the entire collection of instruments page-by-page, or search the instrument database by substantive keywords. The instrument-variable link allows analysts to examine questionnaire skip patterns and item context on-screen, a process which enhances the variable selection process and reduces the need for paper copies of instruments.

Data Extracting. The third feature, Data Extract software, allows the user to produce, with a few keystrokes, SPSS or SAS program statements for any subset of selected variables and then create an active or system file for statistical analysis. This

capability permits analysis of even the largest of data sets to be conducted on most microcomputers. It also saves significant preparation time in writing and re-writing SPSS and SAS program statements to define variables used in a given analysis.

In the following section, the capabilities of these Search & Retrieval, Instrument Link, and Data Extract software programs will be illustrated with a sequence of searches from the AIDS/STD Data and Instrument Archive, which contains over 14,000 variables from 11 major investigations of the incidence and prevalence of specific sexual behaviors, contraceptive use and STD preventive behavior, AIDS/STD knowledge, and attitudes regarding contraception and STD prophylaxis.

Search and Retrieval of Variables from the AIDS/STD Data and Instrument Archive
In the first example, a **search by topic** across the 11 studies comprising the AIDS/STD Archive illustrates how the Search and Retrieval software can be used to find all variables indexed under the substantive Topic of **hiv/aids**. After initiating the program and pressing the **f3** select menu to request a search by topic, the **f4 search** menu is used to perform the search using the wildcard query **topic?** (Figure 1).

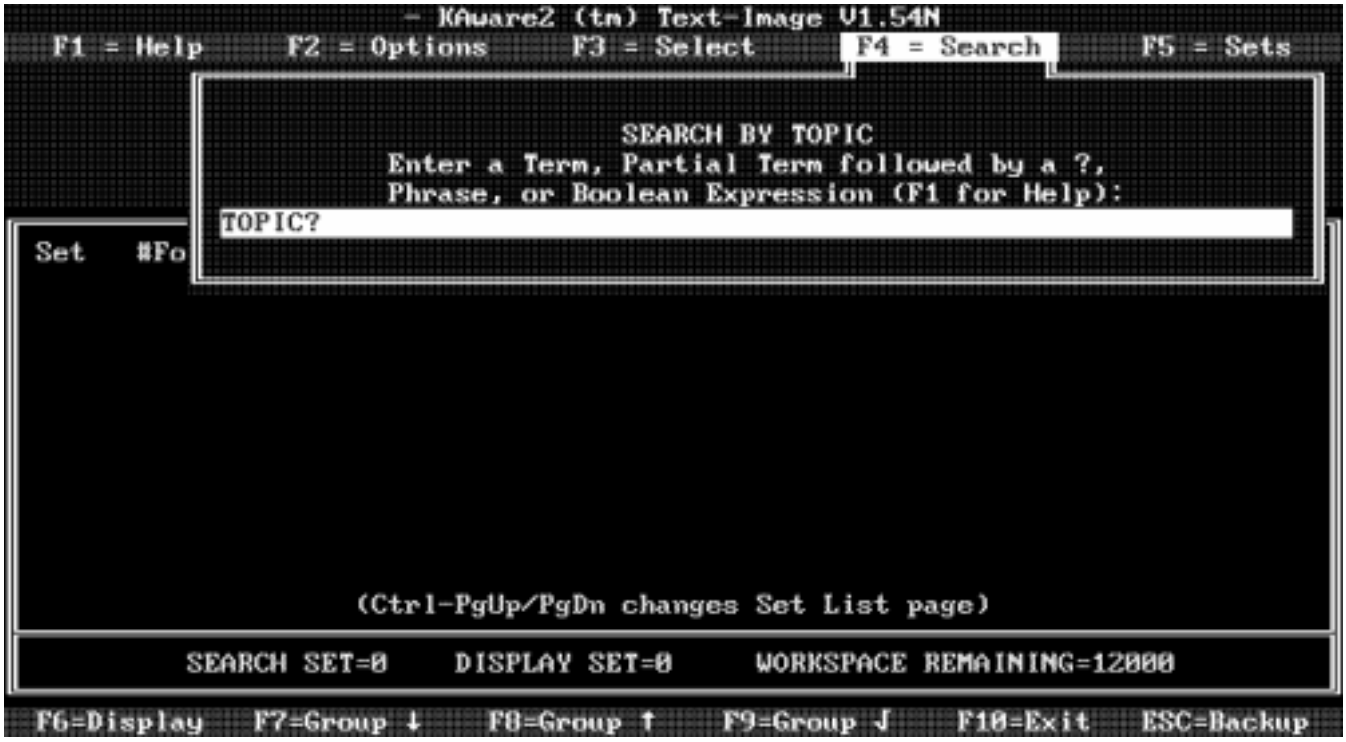


Figure 1

Wildcard searches such as the **topic** search requested in Figure 1 produce a scrollable menu of retrieved items. In this example, the 62 substantive Topics available in the AIDS/STD archive are displayed in the menu. The third of the six screens comprising this menu is shown in Figure 2.

Figure 2 shows that 630 variables in the 11 studies comprising the AIDS/STD archive have been indexed under the topic of **hiv/aids**. Descriptions of each of these "hit" variables can be viewed by highlighting the **hiv/aids** topic line in the scrollable menu and pressing enter. Variables records for all 630 "hit" variables will be retrieved, and the first variable record in the set will be displayed on the screen. Figure 3 shows the first variable record from the **topic = hiv/aids** search set.

As Figure 3 shows, the variable text record provides the variable name and label (Line 1), study name (Line 3), author or investigator names (Line 4), topic and type codes assigned to the variable (Lines 5 and 6), and value labels (Line 7 ff.). The text record also contains the following on-screen instructions (Line 2) for viewing the instrument page containing the original item from which the variable was derived: **"press alt + i to see image of actual questionnaire."** In Figure 4, the instrument page for **hvb01021: a:7a been to hiv testing site** is shown in the upper left corner of the screen. The lower right portion of the Figure 4 screen contains instructions for zoom enlargement of the instrument page image, which can be

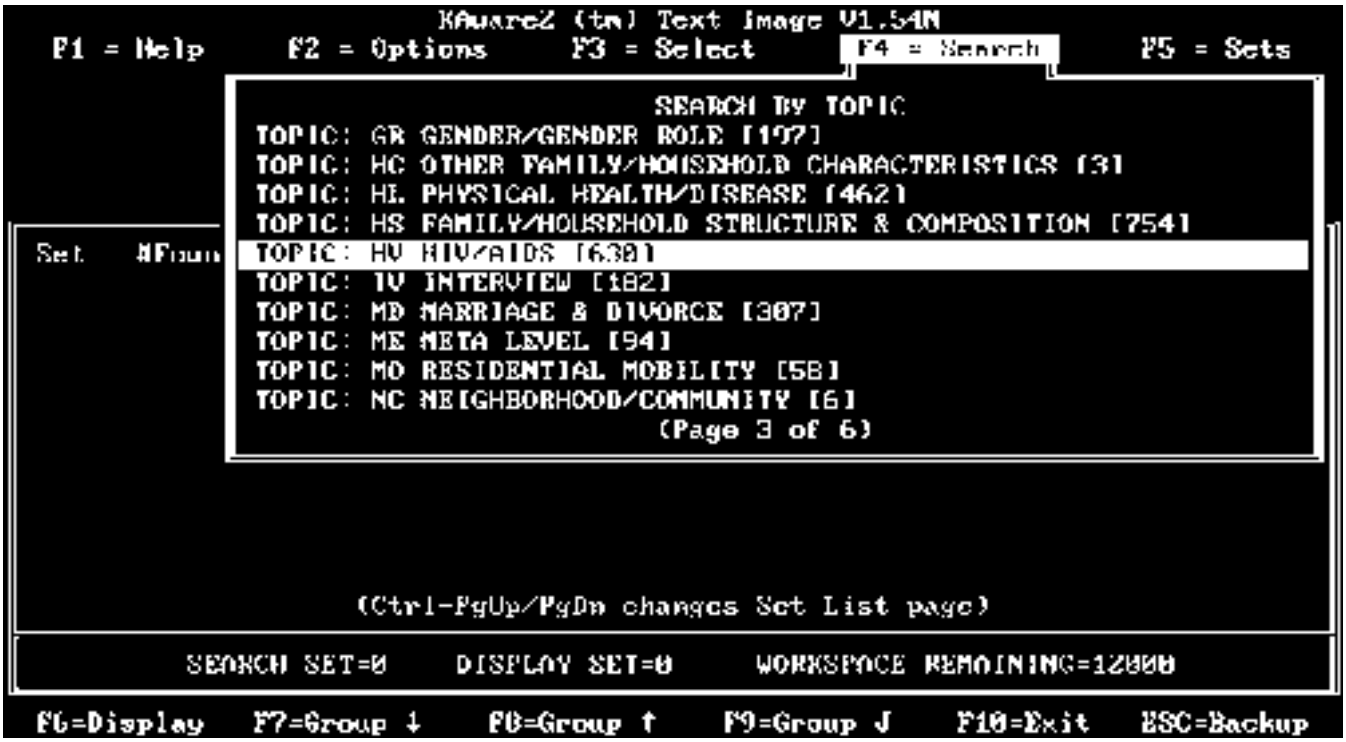


Figure 2

magnified to one of three sizes and printed out. In Figure 5 the instrument page for variable **hvb01021** is magnified at zoom level 2 and has been cropped to fit the page.

To allow the user to quickly locate the original item on the graphic page image, all variable labels include the questionnaire item number. In this example, variable **hvb01021: a:7a been to hiv testing site**, is from item 7.a. on the "A" questionnaire from Data Set 01, *The California Survey of AIDS Knowledge, Attitudes and Behavior: 1987.* Figure 5 shows that item 7.a. is

Figure 3

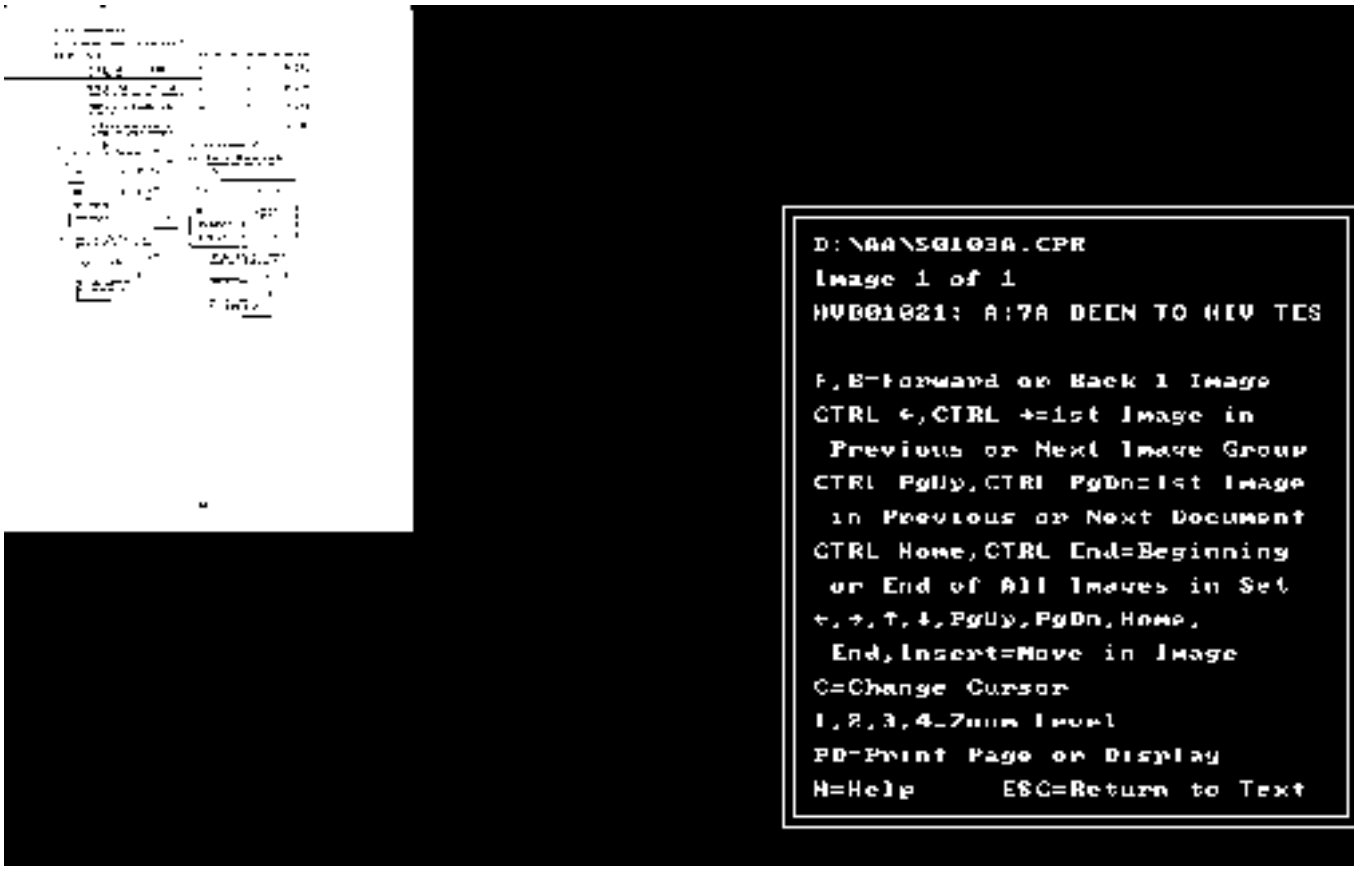


Figure 4

part of a compound question assessing AIDS-related behaviors.

Because one of the primary goals of a search and retrieval session is to select a subset of variables for analysis, the next step in our example will show how a search set of demographic variables including age, race, sex, and community of residence can be combined with the set of 630 **hiv/aids** variables already retrieved, in order to produce a prototypic set of analysis variables. Under the **f4 search by topic** function, the user can select and retrieve multiple topics at one time with the **f9 (group** ÷) key. Our second topic search retrieves all variables in the AIDS/STD Archive that assess age, race or ethnicity, gender or gender role, neighborhood or community, and region or state — a total of 717 variables indexed under these five different topics. Figure 6 shows the fifth of a set of six topic-menu screens, showing how such selection was done for two of the five topics (**race/ethnicity** and **region/state**).

The 717 "hit" variables from this second search covering five demographic topics are combined with 630 hit variables from the first HIV/AIDS-topic search using the Boolean operator **or** (Figure 7).

A new set of 1,347 variables results (Figure 8, "Set 3"). To save the selection of demographic and HIV/AIDS variables comprising Set 3 in a file that can be used by Sociometrics' Data Extract software, the user simply selects the option "Transport a Set" from the **f5 sets** menu and provides a filename for the transported set (Figure 8).

Transforming a Variable Search Set into a Statistical Analysis Package

Command File

The Data Extract software uses the transported set file to create extract command files in user's choice of the SPSS or SAS statistical analysis package. Figure 9 shows the on-screen summary produced after the sample transport file has been read. Each study from the AIDS/STD Data and Instrument Archive contains some of the demographic and HIV/AIDS variables from the search.

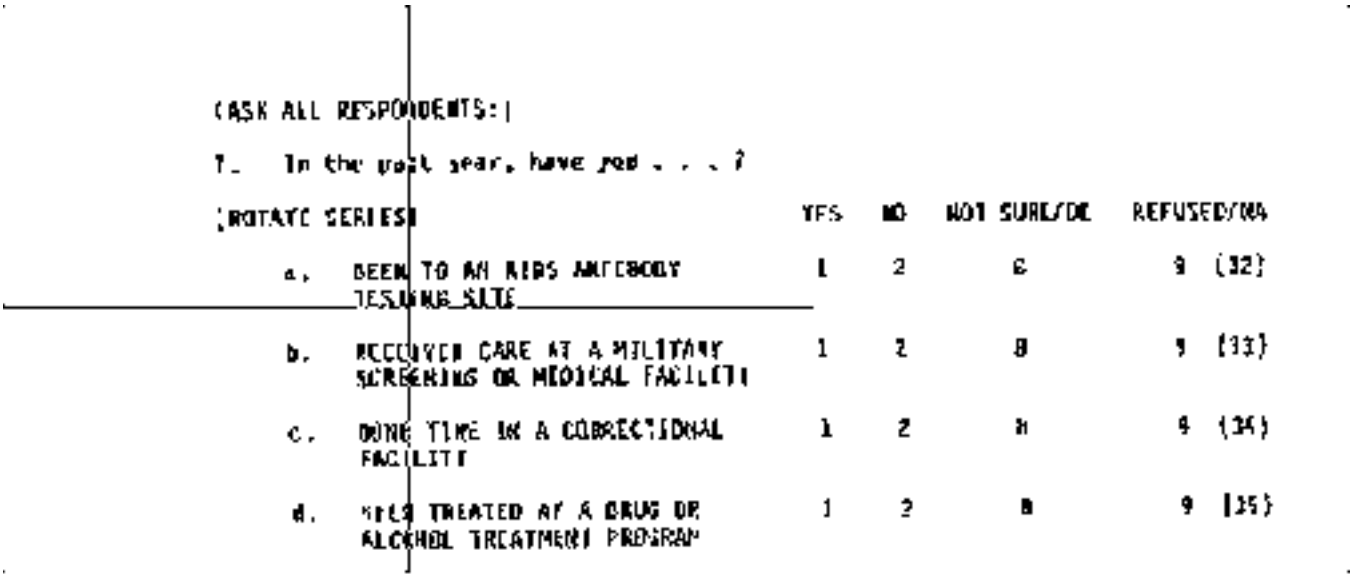


Figure 5

Extract command files may be produced for each data set in turn. The program requires between 30 seconds and 3 minutes to produce an ASCII command file that will create an SPSS/PC+ , SPSSx, or SAS active file with variable names, variable labels, and value labels.

Figure 10 shows a sample extract SPSS-PC+ command file for Data Set 01, the *California Survey of AIDS Knowledge, Attitudes, and Behavior: 1987*. To save space, only a sample of variables from this file is presented; lines edited out of the

SPSS-PC+ command file in Figure 10 are noted with a series of dots (....). The resulting extract command file can be used to

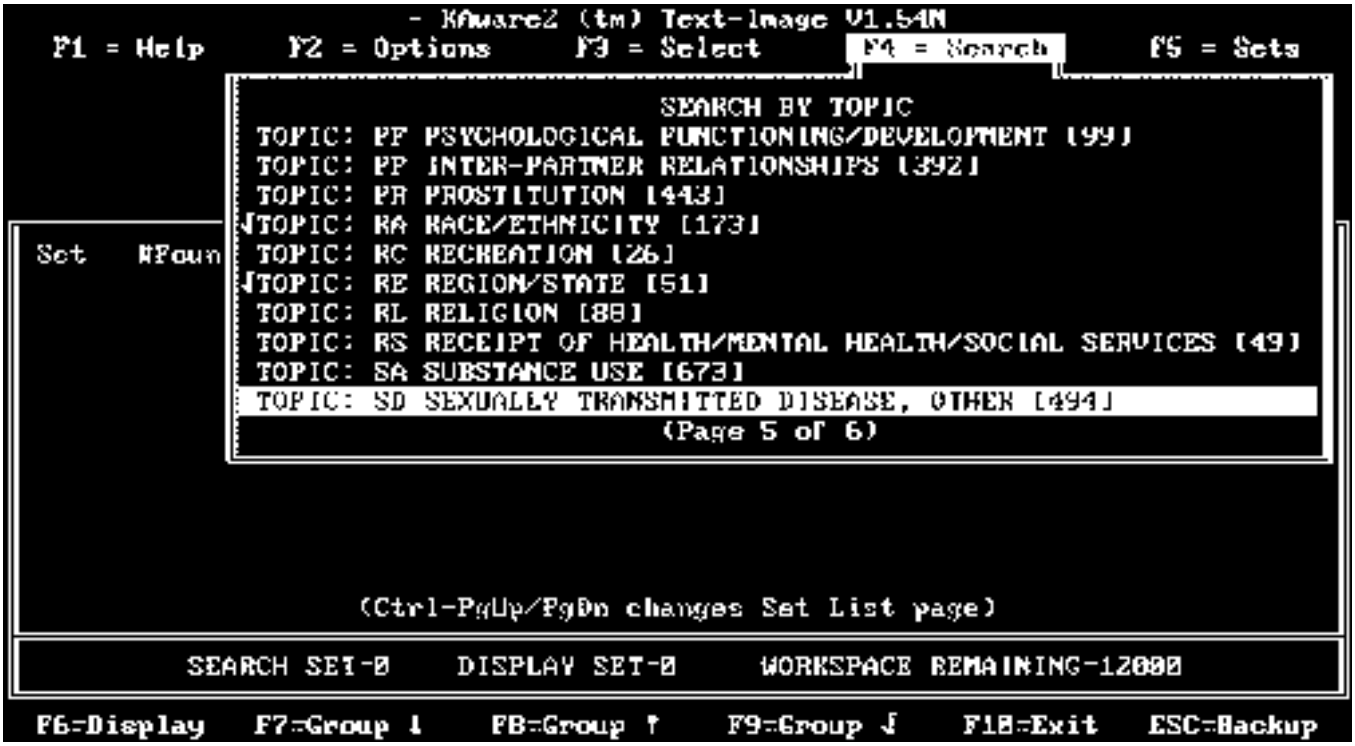


Figure 6

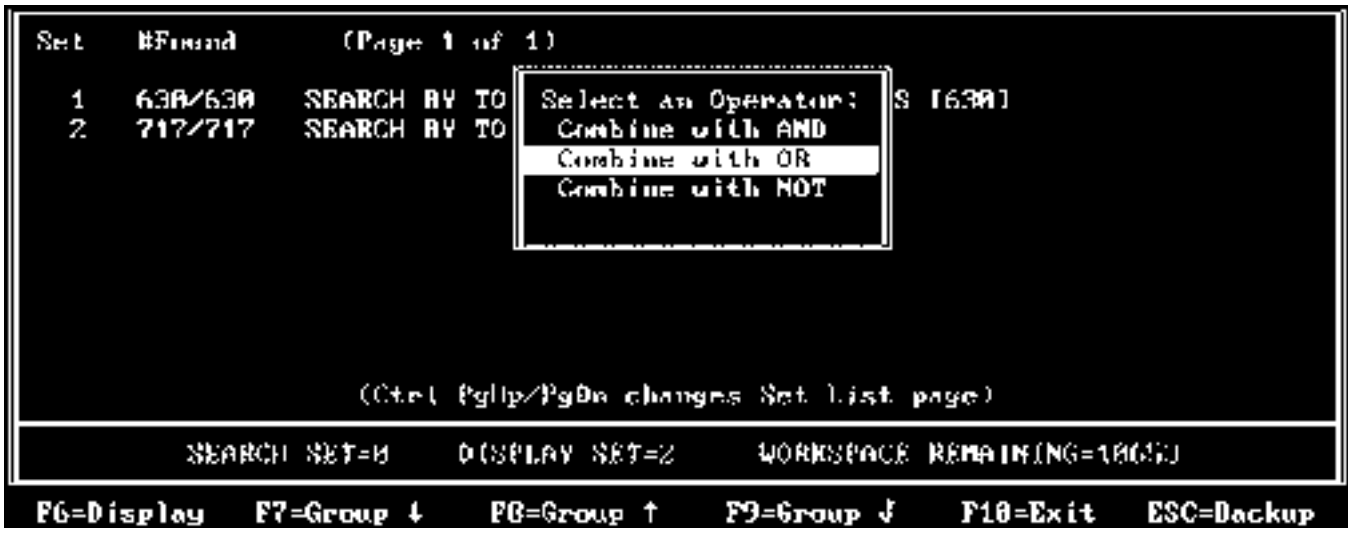


Figure 7

create an SPSS/PC+ active file, or with minor editing, a system file on which analyses of the 100 Age, Race, Gender/Gender Role, Neighborhood/Community, Region/State, and HIV/AIDS variables in the *California Survey of AIDS Knowledge, Attitudes, and Behavior: 1987* can be performed with ease.

Next Step

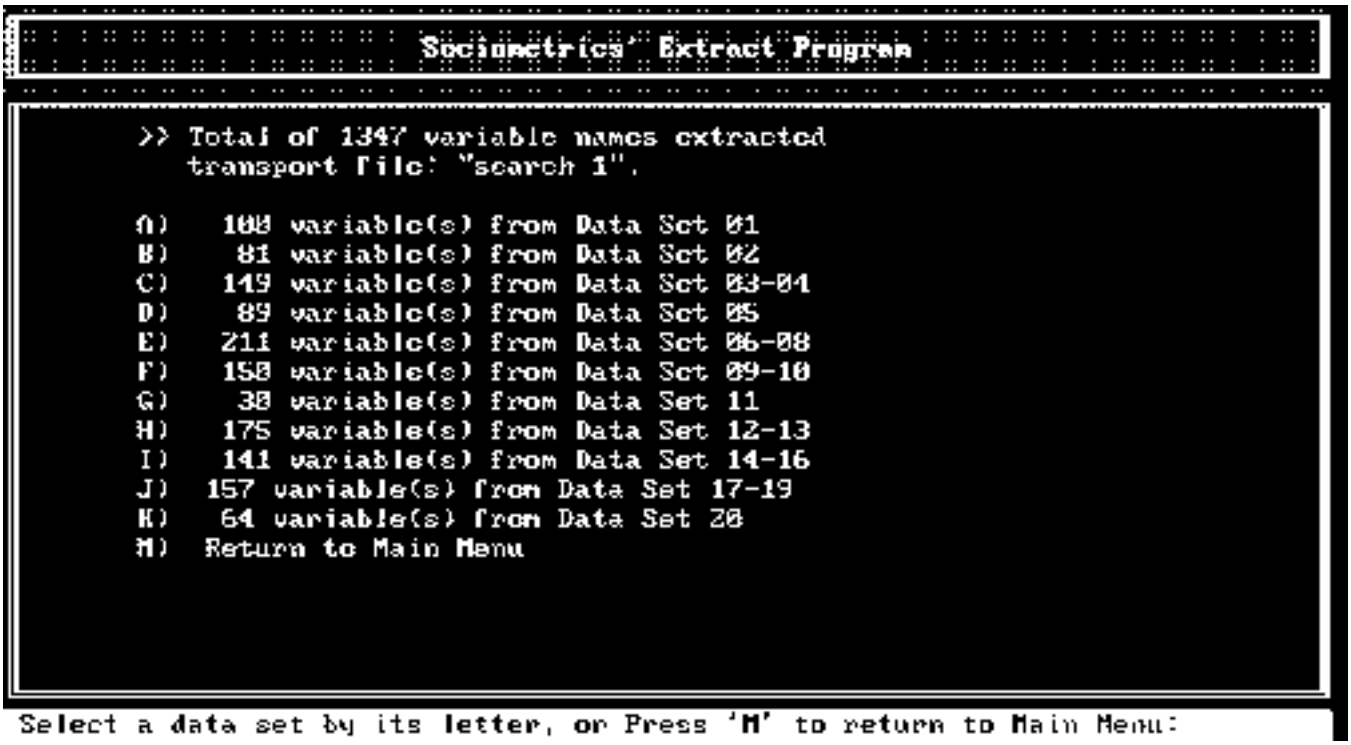
The last decade has witnessed paradigmatic changes in the way social science data sets are stored, delivered to users, and analyzed. Many of these changes have been brought about by rapid technological developments in microcomputer hardware and software, optical storage devices, and the growth of the Internet and on-line digital libraries. This paper has briefly described one state-of-the-art data archive consisting of high quality data in a field of current interest. The AIDS/STD Data and Instrument Archive is representative of all data archives produced at Sociometrics, which combine high quality,

Figure 8

paperless documentation with powerful, yet easy-to-use software for variable search and retrieval, viewing of the original questionnaire page containing the item, and data extraction.

As the need for data sets for secondary analysis grows, social scientists will demand more sophisticated methods of data storage and retrieval. Data archivists will continue to face the challenge of providing users with increasingly well-prepared, paperless, and machine-searchable data collections. In anticipation of user needs, Sociometrics is developing several new enhancements for future data archives. First, programming is being developed for electronic links between variables and descriptive statistics as well as technical notes such as skip logic, scale variables, case weights, and other study-specific information associated with each variable in an archive. Because of the overwhelming amount of information that typically accompanies a data set, it is important that accurate and complete documentation at the level of the individual variable also be accompanied by print-on-demand capabilities. Providing print-on-demand capabilities for user-selected portions of documentation (e.g., user guide sections, questionnaire pages) in a variety of character formats including ASCII, Microsoft Word, WordPerfect and Postscript, constitutes a second forthcoming technological enhancement. Finally, Sociometrics is preparing

Figure 9



to launch SOCIONET, an on-line Data Library service, available 24 hours a day to Internet users.

It is hoped that these developments will move the field of data sharing and secondary data analysis ever closer to the vision of the successful enterprise-of-the-future described by Stanley M. Davis in his book Future Perfect: the ability to meet users' needs—in this case their data information needs—any time, any place, any where (Davis, 1987).

References

Card, J. J. and McKean, E. A. (1993). Harnessing advances in technology: New opportunities for research and teaching in social gerontology. Gerontology & Geriatrics Education, 14,2, 63-76.

Davis, Stanley M. (1987). Future Perfect. Reading, Massachusetts: Addison-Wesley.

1. We wish to thank our colleagues, Drs. Kathryn Muller, Bill Farrell, and Eric Lang, for their comments on an earlier draft of this paper.

```
* Sociometrics Corporation (Los Altos, CA)
* AIDS/STD Data Archive, Data Set 01
* SPSS/PC+ Program Statements for
* The California Survey of AIDS Knowledge,
* Attitudes, and Behavior: 1987.

DATA LIST FILE = 'std01.raw'
 /
 GRS01003  20-27
 /
 /
 HVB01021  20-27
 /
 HVC01031  28-35
 HVC01032  36-43
 HVC01033  44-51
 HVC01034  52-59
 HVC01035  60-67
 HVC01036  68-75
...

VARIABLE LABELS
 GRS01003 "* Respondent sex"
 HVB01021 "A:7a Been to HIV testing site"
 HVC01031 "A:12a AIDS Risk: Use public facilities"
 HVC01032 "A:12b AIDS Risk: Go to school w/PWA child"
 HVC01033 "A:12c AIDS Risk: Donate blood"
 HVC01034 "A:12d AIDS Risk: Live near hospital w/AIDS patients"
....

VALUE LABELS
 GRS01003
       1   "MALE"
       2   "FEMALE"
 /HVB01021
       1   "YES"
       2   "NO"
       8   "NOT SURE/DONT KNOW"
       9   "REFUSED"
....

FINISH.
```