Development and Dissemination of an Electronic Library of Social Science Data

JOSEFINA J. CARD

Sociometrics Corporation

Increases in microcomputer processing speed and hard disk storage space, coupled with decreases in the amount of federal funding available for primary data collection, have made secondary analysis of existing databases an attractive option for research and teaching. Sociometrics has pioneered in making exemplary social science data resources readily available, easy to use, and widely disseminated through the establishment of topically focused data archives in a number of important health and social science areas:

- The Data Archive on Adolescent Pregnancy and Pregnancy Prevention (150 studies comprising 234 data sets and more than 60,000 variables)
- The American Family Data Archive (20 studies comprising 122 data sets and more than 70,000 variables)
- The Data Archive of Social Research on Aging (3 studies comprising 22 data sets and more than 19,000 variables)
- The Maternal Drug Abuse Data Archive (7 studies comprising 13 data sets and more than 5,000 variables)
- The AIDS/STD Data Archive (11 studies comprising 20 data sets and more than 14,000 variables)
- The Research Archive on Disability in the United States (19 studies comprising 40 data sets and more than 23,000 variables)
- The Contextual Data Archive (13 data sets compiled from more than 29 sources and more than 20,000 variables)

DESIGN

A previous article described the bootstrapping process that Sociometrics has successfully employed to advance the field of data sharing in the social sciences (Card, 1996). Each successive data archive has contributed to the substantive advancement of its research field by placing in the public domain the "best-of-the-lot" data in the field. In addition, each successive archive has contributed to the advancement of the data-sharing field by enhancing standards for documentation of public use social science data files (Table 1).

Quality Data

Each data set in the Data Library has been selected for inclusion by a national advisory panel of experts in the topical focus of the archive. Selection has been based on strict scien-

tific criteria of technical quality, substantive utility, policy relevance, and potential for secondary data analysis.

Indexing at the Variable Level

Each variable in each data archive is indexed according to a set of approximately 60 archive-relevant topics that characterize the substance of the variable and approximately 15 types that characterize the kind of measure (e.g., "attitude," "behavior," "status"). This topic and type classification afford users a powerful method of quickly searching for, and then extracting, variables of interest both within and across data sets in an archive.

Quality Documentation

Each data set is made publicly available with a standard set of five machine-readable data and documentation files: File 1, a raw data file; Files 2 and 3, machine-readable SPSS and SAS program statements that fully document the variables and values in the data file; File 4, an SPSS data dictionary; and File 5, SPSS frequencies. Each data set is also accompanied by a printed user's guide (provided in machine-readable form, in addition to printed form, for the more recent archives) composed of a standard set of sections and subsections. The provision of standard machine-readable and printed documentation assists users in familiarizing themselves with the Sociometrics data sets. Once a user has worked with one Sociometrics-packaged data set, it is easy for him or her to work with any of the others. The original instrument and code book are offered as optional, supplementary documentation for each data set, when available. For the more recent archives, the original instrument is distributed in machine-readable form along with the data as a set of graphics files (page images).

Search and Retrieval Software

Powerful search and retrieval software accompanies each data archive. This software allows a user to search an entire topically focused collection, a customized group of data sets created explicitly for a given user, or a single data set; to identify variables of interest across this designated search space; and to save located variables as a search set. Users can conduct (a) full-text keyword searches, including variable names, words in variable labels (question descriptors), and words in value labels (response descriptors); (b) searches by assigned topic and type codes; and (c) searches by study name or assigned data set number. Standard Boolean operators (i.e., "and," "or," "not") can be used to combine search sets.

Linked Images of Original Questionnaire Item and Page

An important innovation achieved by the most recent data archives is the inclusion of linked, electronic images of the original data collection instruments that correspond to the archived data sets. This electronic link between the variables and instruments allows users to obtain a better understanding of actual variable content by viewing, for any variable of interest, the page of the original data collection instrument containing the corresponding item as asked of respondents. The instrument-variable link allows analysts to examine questionnaire skip patterns and item context on screen, a process that enhances the variable selection process and reduces the need for paper copies of instruments. In addition, users also can browse entire original instruments or individual subsections of interest through a feature that organizes the instrument around a topical table of contents.

TABLE 1
Chronological Development of Standard Products for the Sociometrics Data Library

Project or Archive Name	Standard Product
Data Archive on Adolescent Pregnancy and Pregnancy Prevention Sponsor: Office of Population Affairs	Selection of exemplary data sets by a national advisory panel of experts in the field Topic-by-type indexing of variables Machine-readable SPSS program statements Software to search and retrieve variables by topic, type, keyword in variable label, and data set number User's guide (printed)
American Family Data Archive Sponsor: National Institute on Child Health and Human Development	All of the above, plus Machine-readable SAS program statements Software to search and retrieve variables expanded to include keyword search of value labels
Development of Search and Retrieval Software for Archival Data Sponsor: National Science Foundation	Software to create user-designated extracts of data files
Data Archive of Social Research on Aging Sponsor: National Institute on Aging	All of the above, plus Toolkits or tutorials for learning how to use complex data sets Software to search and retrieve variables made available for Macintosh users via SoftPC
Maternal Drug Abuse Data Archive Sponsor: National Institute on Drug Abuse	All of the above, plus Instruments, indexed by section title, included in machine-readable, browsable form User's guide included in machine-readable form
AIDS/STD Data Archive Sponsor: National Institute on Child Health and Human Development	Software to search and retrieve variables expanded to include perusal of instrument page containing item as well as neighboring pages
Research Archive of Disability in the U.S. (RADIUS) and Contextual Data Archive Sponsor: National Institute on Child Health and Human Development	All of the above, plus Topic-by-type indexing of variables expanded to include indexing by secondary topic User's guide (machine readable) included in the search and retrieval space Software to search and retrieve variables expanded to include display and printing of frequencies of retrieved variables

Data Extract Software

Finally, Data Extract software allows users of CD-ROM versions of archived data sets to create customized SPSS or SAS program files containing only those variables of interest to them. This capability permits analyses of subsets of large data sets to be conducted quickly (with rapid turnaround) on most microcomputers. It also saves users significant program

development time writing and rewriting SPSS and SAS program statements to define variables used in a given analysis.

DISSEMINATION

Having achieved what we believe to be a close-to-optimal, cost-effective way to select and prepare data sets for the public domain, we have turned our attention to innovative ways to encourage use of this valuable data resource. The present report focuses on advances in dissemination and user outreach that have taken place over the past 3 years.

A public resource is only beneficial if it is used appropriately. But use cannot occur without potential users being aware of the existence, organization, contents, and capabilities of the resource. Therefore, from the Data Library's inception 15 years ago, we have publicized its contents to individual researchers, professors, and students who could potentially use it. We have used a variety of methods to reach potential users, including distribution of a thrice-yearly newsletter, seeding of a complimentary data catalog, circulation of direct-mail fliers, placement of ads in professional journals, presentations of papers in professional conferences, demonstrations of products at exhibit booths at professional conferences, posting of resource announcements to relevant Internet lists, and publication of papers in relevant scientific journals.

More recent dissemination efforts have turned to three new challenges: first, how to package the entire library of 300+ data sets from seven topically focused collections in a cost-effective fashion; second, how to take advantage of the burgeoning universality of a new technology: the Internet; and, third, how to meet the needs of an important, growing constituency of non–social scientists: data librarians.

In talking to our customers, we discovered that what end users—researchers, professors, and students—appreciate the most is quick access to high-quality data. In contrast, librarians are primarily concerned with archival preservation of these important resources. To meet the differing needs of both these constituencies, we created a new package consisting of all the data sets from all of our data archives, the Sociometrics Social Science Electronic Data Library (SSEDL). We put together three versions of SSEDL: two Internet versions and a CD-ROM version.

Our Internet server (www.socio.com) hosts a couple of SSEDL suites. The first suite allows all Internet users to download SSEDL's data sets on provision of a credit card cybercash payment. The second Internet suite allows faculty members and students of SSEDL Data Consortium member institutions to download SSEDL's data sets for free. Membership in the SSEDL Data Consortium is obtained by the institution's library purchasing the CD-ROM version of SSEDL for a fraction of what the data sets would have cost separately (less than $10, as opposed to $225 per data set). This way, both the end user's need for quick access to high-quality data and the librarian's need for preservation of the same data are simultaneously met in cost-effective fashion.

We have supplemented our ongoing dissemination efforts with several innovative ways of reaching our target constituencies. First, we have begun teaming with professional associations of social scientists and librarians to codisseminate SSEDL to their members at a discounted price. Second, we have developed multimedia descriptions and demonstrations of SSEDL, both on CD-ROM and on our web site (http://www.socio.com/edl.htm). Third, we are offering members of the SSEDL Data Consortium an opportunity to keep their collection up to date by means of low-cost subscriptions to SSEDL. Subscribers are provided with annual updates to the collection on CD-ROM as well as ongoing access to the free data set download area of the SSEDL Internet suite.

PEERING INTO THE FUTURE

We will continue expanding the content and capabilities of our data set collections. We will continue the vigorous dissemination of this valuable resource both through our direct efforts and through collaborations with professional associations of scientists and librarians.

We are currently developing two products related to SSEDL. BSRI, the Behavioral Science Research Instruments Archive, will contain searchable, edit-ready, and print-ready machine-readable versions of the demographic, behavioral, and health science instruments—questionnaires, medical forms, interview protocols—used to collect the data in SSEDL. Norms for the scales and items comprising the BSRI instruments will be included in the archive in the form of scale means and standard deviations, item frequencies or response distributions, and item cross tabulations with age, race/ethnicity, and gender obtained from the linked SSEDL data archives. BSRI will also contain a link to the corresponding SSEDL data file that will allow a researcher to select variables for a fully documented SPSS or SAS analysis extract file from BSRI's variable listing, original instrument, and/or item statistics.

MIDAS, the Multivariate Interactive Data Analysis System, will allow online analysis of the data in SSEDL. Online data analytic procedures will include weighted and unweighted frequencies, percentiles, and measures of dispersion and central tendency, as well as two-way and n-way tables with measures of association, comparison of means (two-group and ANOVA) and correlations, and the calculation of complex variance estimations. Users will be able to define case subsets, recodes, or aggregations for analysis and then produce output that can be downloaded or printed. Custom data set download will also be available.

REFERENCE

Card, J. J. (1996). Development of the Sociometrics Data Library on Families, Aging, Substance Abuse, and AIDS. *Social Science Computer Review, 14*, 305-309.

Josefina J. Card, Ph.D., is the president of Sociometrics Corporation. Comments or questions can be addressed to her by mail: Sociometrics Corporation, 170 State Street, Suite 260, Los Altos, CA 94022; by telephone: 650-949-3282, ext. 211; or by e-mail: jjcard@socio.com.