REPORTS AND COMMUNICATION

Development of the Sociometrics Data Library on Families, Aging, Substance Abuse, and AIDS

JOSEFINA J. CARD
Sociometrics Corporation

A number of challenges face the preparer-provider of social science data for public use. The first and second are the provision of data and documentation of the highest quality. The third is the coupling of such data and documentation with powerful yet user-friendly ways to search through the collection, and find data sets and variables of relevance to the research or policy problem at hand. The fourth challenge is to include data extract capabilities, so that smaller, equally well-documented data files containing only variables or cases of interest to the user can be created on demand (thus lowering hard-disk-storage and processing-time burdens). The fifth challenge, dissemination to potential users of the availability of the data resource, is also key; scientists cannot use what they do not know exists. Finally, easy and low-cost access to the data and documentation must be provided.

This report describes the procedures and products of one data preparer-provider, Sociometrics Corporation, to illustrate how one data archive center has addressed these challenges. The Sociometrics Data Library has been in operation for 12 years. Many aspects of the Data Library's operation are generic and can be adapted to other data archives covering other substantive fields; these generalizable aspects of the approach and results merit highlighting and scrutiny.

THE SOCIOMETRICS DATA LIBRARY

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. Table 1 presents an overview of the development of the five topically focused data archives composing the current Sociometrics Data Library: the Data Archive on Adolescent Pregnancy and Pregnancy Prevention (129 studies comprising 196 data sets and more than 70,000 variables), the American Family Data Archive (14 studies comprising 36 data sets and more than 20,000 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

AUTHOR'S NOTE: Josefina J. Card, Ph.D., is the president of Sociometrics Corporation. Comments, questions, or requests for abstracts of the data sets composing the Sociometrics Data Library may be addressed to her by mail: Sociometrics Corporation, 170 State Street, Suite 260, Los Altos, CA 94022; by phone: 415-949-3282; or by e-mail: xb.h330forsythe.stanford.edu.

Social Science Computer Review, Vol. 14 No. 3, Fall 1996 305-309
© 1996 Sage Publications, Inc.

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

Project or Archive Name	*Standard Products*
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 (*DOS* platform) 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

(7 studies comprising 13 data sets and more than 5,000 variables), and the AIDS/STD Data Archive (12 studies comprising 20 data sets and more than 16,000 variables). Two other data archives are under development: the Research Archive on Disability in the United States (RADIUS) and the Contextual Data Archive.

As Table 1 shows, a bootstrapping process has been used in pursuit of the previously described data preparation challenges. Each successive archive has contributed to the substantive advancement of its research field, by placing in the public domain the "best-of-the-lot" data in its field. In addition, each successive archive has contributed to the advancement of the data sharing field by developing new documentation, search, and data extraction procedures and products that can be added to future data sets at little or no cost. How has the Sociometrics Data library addressed the six challenges facing the data preparer-provider?

CHALLENGE 1: QUALITY DATA

The Data Library has solicited and obtained the cooperation of eminent scientists working in the fields addressed by each of its topically focused collections. Each of the data sets 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 accomplished using strict scientific

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

CHALLENGE 2: QUALITY DOCUMENTATION

The various topically focused collections in the Data Library have been formed over time with grant and contract funding from various agencies of the federal government. Each successive data archive has included all the standard products of the preceding archive and added methodological and technological contributions of its own.

Variable indexing. Each variable in each data archive is coded according to a set of approximately 60 archive-relevant topics that characterize the substance of the variable (e.g., for the Maternal Drug Abuse Data Archive: "Cocaine," "PCP," "Neurobehavioral Function/ Development") and approximately 15 types that characterize the kind of measure (e.g., "Attitude," "Behavior," "Status"). This topic and type classification affords users a powerful method of quickly searching for and extracting variables of interest within and across data sets.

Standard machine-readable files. Each data set is made publicly available with a standard set of five machine-readable data and documentation files: a raw data file (File 1); machine-readable SPSS and SAS program statements that fully document the variables and values in the data file (Files 2 and 3); an SPSS data dictionary (File 4); and SPSS frequencies (File 5).

Standard user's guide. Each data set is made publicly available with 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.

Supplementary documentation. The original instrument and codebook are offered for each data set. 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).

CHALLENGE 3: POWERFUL SEARCH CAPABILITIES

Search and retrieval capabilities. Powerful search and retrieval software accompanies CD-ROM versions of archived data sets. This software allows a user to search an entire topically focused archive, a customized collection 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.

Development of electronic link between data and instrument. An important innovation achieved in the most recent archives (the Maternal Drug Abuse and AIDS/STD 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 can also browse entire original instruments or individual subsections of interest through a feature that organizes the instrument around a topical table of contents.

CHALLENGE 4: DATA EXTRACT CAPABILITIES

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.

CHALLENGE 5: VIGOROUS DISSEMINATION

The existence, contents, and products of the Sociometrics Data Library are disseminated with equal vigor to individual end users (scientists, educators, students) and to institutions (libraries, academic departments) using a variety of methods. These methods include a semiannual newsletter, direct mail fliers, ads in professional journals, presentations of papers in professional conferences, demonstrations of products at exhibit booths at professional conferences, resource announcements to relevant Internet lists, and publication of professional papers.

CHALLENGE 6: EASY, LOW-COST ACCESS

A catalog for each data archive describes each study in the collection with its data set number(s), title, principal investigator(s), study abstract, available products, and prices. These catalogs are available free of charge upon request. All products can be ordered directly from Sociometrics by mail, phone, or fax. Turnaround time between receipt of order and shipment is 7 days or less. Prices for all data and documentation products are set to recover the cost of dissemination, production, and lifetime technical assistance on effective use. Substantial discounts are given to individuals, libraries, or academic departments that purchase an entire topically focused data archive collection.

PEERING INTO THE FUTURE

Archive-related development continues at Sociometrics, under the aegis of RADIUS, the American Family Data Archive, Volume II, and the Contextual Data Archive. Included in this ongoing development are the expansion of indexing of each variable from one to two topics (if needed) and the expansion of the search and retrieval software accompanying each archive to include (a) textual searches of all Sociometrics-prepared user's guides; (b) display and printing of unweighted frequencies for all "retrieved" or "hit" variables; and (c) direct

and efficient performance of the search and retrieval and data extract software on *DOS*, *Windows*, Macintosh, and *UNIX* platforms. Under a grant from the National Institute of Child Health and Human Development, Sociometrics is also establishing SOCIONET, an on-line digital library of exemplary social science data. SOCIONET will make all of the Sociometrics Data Library's data, documentation, and software available to the nation on-line, 24 hours per day, via the Internet.

These developments address several of the previously described challenges to the data archivist, carrying the field of public-use data provision further toward quality documentation, coupled with powerful search techniques, to help users find what they need, when they need it.