User-Centered Design and Innovation in the Sociometrics Social Science Electronic Data Library (SSEDL)

*by Josefina J. Card, Tamara Kuhn, and Thomas Wells**

Abstract

This paper presents the current state (scientific content, formats, platforms, distribution partners) of the Sociometrics Data Archives, collectively known as SSEDL, the Social Science Electronic Data Library. It then peers into the future by describing areas of topical expansion, new target audiences, and new science-based resources currently being built around SSEDL. Usage information is also given.

In the twenty years since the first topically focused data archive was established at Sociometrics (Card 1989, Card 1996, Card 2000, Carley and Card, 2000), there has been a tremendous increase in the availability of inexpensive and powerful computing resources (Davey et al., 2006), an expansion of federal requirements and incentives for data sharing (Melichar, Evans, and Bachrach, 2002, NIH, 2003), and a burgeoning of cost-effective data distribution options such as CD-ROM and the Internet. These factors have made use of data archives an increasingly attractive option for research and teaching, and have spurred the development of diverse collections of primary research data for conducting secondary research.

The data archives at the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan are the largest collection of social and behavioral research data. The data archives at Sociometrics - collectively known as SSEDL, the Social Science Electronic Data Library - continue to be an attractive supplement, especially for users interested in public health issues and in use of data for novice researchers or for teaching purposes. The continual addition of new datasets to SSEDL makes the resource a rich source of data for those in the public health, medical, nursing, social work, and social science professions. In this article we provide an overview of the current content of SSEDL. We describe the features that make SSEDL easy to use by novice and expert researchers alike. We end with a look into the future and share plans for upcoming content and user-focused innovations.

Organization and Content

The Sociometrics Social Science Electronic Data Library is a premier health and social science resource that is comprised of nine topically-focused data archives. Each data archive has exemplary datasets selected by a distinguished Scientist Expert Panel for their scientific merit, substantive utility, potential for secondary data analysis, and program or policy relevance. Table 1 gives an overview of the contents of SSEDL. Details on each archive, including a complete list and description of included datasets, can be found at www.socio.com/dataarchives.htm. With some 600 datasets from more than 250 different studies comprising nine topically-focused collections, SSEDL is a unique source of high quality health and social science data and documentation for researchers, educators, students, and policy analysts. More than eighty percent of the SSEDL collection is unique and not available from any other public source (including ICPSR).

Table 1: Overview of Sociometrics' Data Archive Collection

Topically-Focused Archive	Studies	Datasets	Variables
Adolescent Pregnancy	162	286	80,000
Aging	3	22	19,000
Child Well-Being & Poverty	12	36	20,000
Complementary & Alternative Medicine	8	17	10,000
Contextual	13	29	19,000
Disability	19	40	25,000
Family	20	122	66,000
HIV / AIDS / STD	19	30	19,000
Maternal Drug Abuse	7	13	5,000
TOTAL	**263**	**595**	**263,000**

User-Focused Features

Product Packaging
Each dataset in the collection is made available with a standard set of eight machine-readable data and documentation files: (1) the raw data file, (2) SPSS program statements that define each variable in the dataset and provide both variable and value labels, (3) SAS program statements, (4) SPSS data dictionary, (5) SPSS frequencies, (6) an SPSS portable file, (7) a SAS transport file, and (8) a User's Guide with standard sections: description of study, description of machine-readable files, complete list of variables sorted by their topic and type, frequencies for key variables included in most datasets (e.g., race, gender, marital status, etc.), and results of data completeness and consistency checks conducted by archive staff. This standard packaging and documenting of each dataset in SSEDL assists users in familiarizing themselves with the resource. Once a data analyst has worked with one SSEDL dataset, it is easy for him or her to work with any of the others in the collection.

Search Aids
Data users are able to identify datasets and variables that meet their needs and specific variables of interest via a search mechanism freely available on Sociometrics' web site (www.socio.com/search.htm). Analysts can specify whether they want to search the entire SSEDL collection, a combination of data archives, or a single data archive. The keyword search utilizes standard Boolean search strings and searches key fields that include variable labels, value labels, study name, and investigator names. For each variable that the search returns, the display shows the variable label, the value labels, the names of the original investigators, and the study name with a link to additional study information (brief abstract, summary of methodology, number of variables, number of cases, and purchase options).

Product Formats
The format of data distribution has changed significantly over the past twenty years, in keeping with technological advances in this period. Initially, large datasets were made available on mainframe tape and smaller datasets were made available on diskette. Now datasets and accompanying documentation are distributed in user's choice of CD-ROM or Internet download.

Multi-Level Acquisition Options
Purchasers can acquire data in one of three configurations: an individual dataset, a complete topical archive, or the complete SSEDL collection (currently nine topical archives).

Individual datasets can be obtained on CD-ROM or downloaded from the Sociometrics web site.

Complete topical archives can be ordered on CD-ROM at a cost that is significantly less than purchasing each of the datasets individually.

The complete SSEDL collection is available to universities and other institutions via subscription through Thomson Gale, the exclusive worldwide distributor of SSEDL. All faculty, staff, and students at the subscribing institution are allowed free and unlimited access (via Internet download) to all of the several hundred SSEDL datasets and data-related materials. Additionally, subscribers have immediate download access to new datasets as they become available. This dissemination format meets users' need for quick access to the data, relieves the burden on data librarians of providing access to the data, and is extraordinarily cost effective.

Usage Report
During the past five years, the explosive growth of the Internet is reflected in the increase in usage rates of Sociometrics' web site and data-related web features. As seen in Table 2, during the past five years the number of downloads of data-related products, including datasets, has increased by nearly 300% and the number of "hits" to data

Table 2. Internet Usage & Purchase Rates of Sociometrics' Data Archive Collection, by Year

	Year				
	2001	**2002**	**2003**	**2004**	**2005**
Visits to Sociometrics website	136,598	145,237	175,984	182,528	253,636
Hits to all data archive pages	42,003	70,022	86,770	101,819	210,494
Downloads of data and data-related products	10,704	16,325	29,729	34,954	39,672
Number of Units Ordered (non-subscribers	171	222	156	167	193
Number of Purchasers (non-subscribers	82	105	91	98	138

archive-related web pages has increased by nearly 400%. In the past year alone, the number of hits to all data archive pages has more than doubled. The large increases in hits and visits appear to be a function of an increase in referrals from search engines, with the majority of new visitors being referred from Google, Yahoo, and AOL.

As more datasets are added to the archive collections and as Internet use continues to increase we anticipate the rates of both web site visits and dataset usage to continue to increase in the coming years.

Looking toward the Future: Expanding Data Archives and New Technologies

We are continually expanding the content and capabilities of our data archives. As we move toward the future, user-focused innovation will be present in both content and the technology accompanying selection and use of the data. We are in the process of adding three new topically-focused data archives. The first is the Data Archive of Longitudinal Studies on Childhood Problem Behaviors, which is being established with funding from the U.S. National Institute of Mental Health. This archive will consist of an online and CD-ROM collection of important longitudinal studies on childhood problem behaviors. The archive will include content-rich longitudinal studies that are not currently available for public use, as well as those studies in the public domain that have not received widespread use

Complementary and Alternative Medicine Data Archive
Topic and Type Distribution

Mouseover any topic or type to view a definition.	Attitudes /Values	Behavior	Clinical Diagnosis	Cognition	Emotion	History	L
Acupuncture	478	69	280	1	84	47	
Age						20	
Alternative Medical Systems, Other		4				4	
Biological Function, Reproduction		22	36			65	
CAM Therapies, General		17				16	
Chiropractic		6				4	
Conventional Therapies		224		1			
Psychological Function, Development		420		12	66		
Quality of Life, General Health							
Race, Ethnicity							
Reflexology		4				3	
Region, State							
Religion							
Supplements, Vitamins		1					
Traditional Chinese Medicine		5					
Wealth, Finances		8				4	
Total	553	3317	395	15	217	572	

Figure 1. Topic and Type Distribution Search Matrix Interface (partila view)

among the research community. The second new archive, the Communication Disorders Data Archive, is being established with funding from the U.S. National Institute on Deafness and Other Communication Disorders. This archive will house state-of-the-art research datasets that address the prevalence and the social, behavioral, and occupational antecedents and consequences of hearing impairment and speech and language disorders. Both archives have recently completed the dataset selection stage and each has an archived dataset. The objective of our newest archive, the Welfare Reform Evaluation Data Archive, is to facilitate access to high quality welfare reform evaluation studies that will enable welfare policy research among a broad pool of scholars and researchers.

Updates to Current Archives

In addition to creation of new topically-focused data archives, the existing archives in the Social Science Electronic Data Library are continually being expanded through the addition of new datasets. The U.S. National Institute of Child Health and Human Development is providing funds for the addition of datasets each year to the Data Archive on Adolescent Pregnancy and Pregnancy Prevention (DAAPPP). Recently archived DAAPPP datasets include the National Longitudinal Study of

Sociometrics Corporation

Search Results

Your Query "**md meditation AND b behavior**" matched 5 documents out of 9930.
5 documents displayed.

0.80 CAM 04-05 Variable: MDB04861 F12L: Past year: relaxation techniques fo:

12-Jan-2005 04:55:49 pm,
http://www.socio.com/srch/variable/camda/cam0405/CAM04861.HTM
Excerpt (section from the web page containing the hit phrase):
Archive Name : Complementary and Alternative Medicine Data Archive (C.
National Survey of Self-Care and Aging (NSSCA), 1990-1994 Investigator
Jean E. Kincade Norburn Data Set No(s). : CAM 04-05 Variable Name : M

0.80 CAM 17 Variable: MDB17097 Ever seen hypnotherapist

12-Jan-2005 05:06:07 pm, http://www.socio.com/srch/variable/camda/c
Excerpt (section from the web page containing the hit phrase):
Archive Name : Complementary and Alternative Medicine Data Archive (C.
Use and Expenditure on Complementary Medicine in England: A Population
Investigator(s) : Kate Thomas, Jon Nicholl, Patricia Coleman, Christian Stac

Figure 2. Topic and Type Distribution Search Results

Adolescent Health (Add Health), Wave III, 2001-2002, the Public Use Education Data; National Survey of Family Growth, Cycle 6, 2002; and the National Longitudinal Study of Adolescent Health, Wave III, 2001-2002 (Add Health). The other archives shown in Table 1 are in the process of being updated and prospective datasets for each archive are currently being prepared for review by a Scientist Expert Panel. At the conclusion of this cycle, each archive will have been updated with new datasets.

An Upcoming User-Focused Innovation: Guided Search through a Data Archive's Topical "Areas of Richness"

A key element of assisting researchers in the use of secondary data is helping them identify the best datasets for their research questions and topics of interest. Although users can currently perform web-based keyword searches on variables in each of Sociometrics' archives, it was determined that the search process could be made more productive if users were given a broader overall sense of the areas of topical areas of richness within each of the archives, and then were able to identify specific variables of interest within those topics. As a result we have begun development of a simple, cost-effective search interface to meet that need.

[Next Doc]

Archive Name : Complementary and Alternative Medicine Data Archive (CAMDA)

Study Name : National Survey of Self-Care and Aging (NSSCA), 1990-1994
Investigator(s) : Gordon H. DeFriese, Jean E. Kincade Norburn
Data Set No(s). : CAM 04-05

Variable Name : MDB04861
Variable Label : F12L: Past year: relaxation techniques for pain
Topic 1 : MD MEDITATION, YOGA, RELAXATION
Type : B BEHAVIOR

Value Label(s) :

Value	Value Labels
0	No
1	Yes

[Next Doc]

Figure 3. Variable Level Search Output

The search interface for each data archive displays matrices of variables by the topic and type distribution for that archive. Figure 1 shows the prototype matrix for this new search capability. In Figure 1 the "areas of richness" of the Complementary and Alternative Medicine Data Archive can be seen from the cells with high numbers (many variables of the given topic and type). Using a small JavaScript program that generates a help balloon when called by a mouse-over, each topic and type is clearly defined for the user by simply placing the mouse pointer over the topic and type heading in the matrix. The number of variables in the archive associated with any topic and any corresponding type are displayed in the matrix. Each of the numbers in the matrix is a link that calls a pre-populated defined keyword query to the Verity system requesting a search for variables containing only the topic and type corresponding to that box of the matrix.

For example, clicking on the number 5 in the cell corresponding to the TOPIC = Meditation, Yoga, and Relaxation and TYPE = Behavior yields the referenced five "hit" variables. Figure 2 gives the first screenful (four) of these variables. As seen in Figure 2, the search returns a formatted list of variables matching the search topic and type keyword search criteria. Each variable in the result list is displayed with the variable name, variable label, and an excerpt from the HTML page that corresponds to that variable's information within the search index.

Clicking on a hit variable's name and label then returns a web page displaying the variable name, variable label, value labels, the variable's topic and type codes, and information about the dataset including the study title and the original investigators. For example, clicking on the first "hit" variable in Figure 2, "Past Year Relaxation Techniques for Pain," results in information on the metadata associated with this variable (Figure 3).

Finally, clicking on the study title returns complete information about the study and provides links to purchase or download the dataset, if desired.

This user-focused search mimics the thinking of the analyst in searching for data that might address his or her topic of concern. First the analyst is advised in advance of the "areas of richness" of a data archive (Figure 1). Then s/he is systematically guided through the contents of the data archive, through the topics and types by which all of the several hundred thousand variables in SSEDL have been indexed. Finally, metadata about the variable and acquisition information about the dataset are provided.

Conclusion

During the past two decades the evolution of Sociometrics' data archives has reflected current trends and changes in technology, in this manner expanding the definition and potential usage of data archives. During the next decades we plan to continue the expansion of the collections' content and capabilities and continue the focus on data quality and user-focused design, search, and dissemination.

* This article was presented at the IASSIST 2006 conference in Ann Arbor, Michigan, at the session "Innovations in Data Dissemination". The authors Josefina J. Card, Tamara Kuhn, and Thomas Wells are all at Sociometrics Corporation. Correspondence to: Dr. Josefina J. Card, Sociometrics Corporation, 170 State Street, Suite 260, Los Altos, CA 94022, (650) 949-3282 x211, jjcard@socio.com.

References

Card, J.J. (1989, January). Facilitating data sharing. ASA Footnotes.

Card, J.J. (1996). Development of the Sociometrics data library on families, aging, substance abuse, and AIDS. Social Science Computer Review, 14, 305-309.

Card, J.J. (2000). Development and dissemination of an electronic library of exemplary social science data. Social Science Computer Review, 18, 82-86.

Carley, M. and Card, J.J. (2000). The Social Science Electronic Data Library: Serving the needs of data librarians and users. IASSIST Quarterly, 24, 8-14.

Davey, M.E., Matthews, C.M., Moteff, J.D., Morgan, D., Schacht, W., Smith, P.W., Morrissey, W.A. (2006). Federal research and development funding: FY2007. Congressional Research Service Report. Library of Congress.

Melichar, L., Evans, J., & Bachrach, C. (2002). Data access and archiving: Options for the Demographic and Behavioral Sciences Branch. National Institute of Child Health and Human Development (NICHD), August.

National Institutes of Health. (2003). NIH Data Sharing Policy and Implementation Guidance. http://grants1.nih.gov/grants/policy/data_sharing/