The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors
The QUARTERLY is published four times per year. Authors are encouraged to submit papers as word processing files. Hard copy submissions may be required in some instances. Manuscripts should be sent to Editor: Karsten Boye Rasmussen.

The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press Manual of Style or Kate L. Turabian's Manual for Writers. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines", Journal of the American Society for Information Science 30(2):77-82, March 1979. Announcements of conferences, training sessions, or the like, are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event.

Editor
Karsten Boye Rasmussen
Department of Marketing & Management
University of Southern Denmark, SDU
Campusvej 55, DK-5230
Odense M, Denmark
Phone: +45 6550 2115
Email:kbr@sam.sdu.dk

Production
Fadi Dagher
Minnesota Population Center, Columbia University
537 Heller Hall
271 19th Avenue South.
Minneapolis, MN 55455.
Phone: 612-624-7091
Email:fd2102@columbia.edu

Walter Piovesan
Maps/Data/GIS Library
Simon Fraser University
Burnaby, B.C.
Canada V5A 1S6.
Phone: (604) 291-5869.
Email:walter@sfu.ca

Title: Newsletter - International Association for Social Science Information Service and Technology
ISSN - United States: 0739-1137 © 1997 by IASSIST. All rights reserved.
Welcome to the second issue of the IASSIST Quarterly, vol. 30.

Winter came late in Denmark, but suddenly the situation was normal: snow came and traffic stopped. And now 14 days later spring is here. I sat outside in the sun reading articles. Still it’s more difficult to write articles outside; we are waiting for improvements to the computer screens. However, some people have stayed inside to write articles for the IASSIST Quarterly. They are presented below.

The first article, “Microdata Information System - MISSY,” is written by Andrea Janssen and Jeanette Bohr from Centre for Survey Research and Methodology, ZUMA, at Mannheim in Germany. The MISSY was presented by Andrea Janssen, Jeanette Bohr, and Joachim Wackerow at the session on “Effective Strategies for Metadata Management” at the IASSIST conference in May 2006 in Ann Arbor. The data in the system are from the German microcensuses for 1995 and 1997, which contain a sample of one percent of all German households. The microcensus has been carried out since 1957, and parts of the microcensus are available for research. The researchers need extensive metadata on both the study and variable level, e.g., the microcensus uses complicated classifications of professions, sectors, and household arrangements. The system is based on the standard from the Data Documentation Initiative and documentation includes general information such as questionnaire, codebooks, interviewer guides, frequencies and also some tips and recommendations on use of the data. MISSY is a German system, in German language, and for researchers in Germany. However, all are free to gain from the experiences presented in the article.

The second article was also presented at the IASSIST 2006 conference, at the session “Innovations in Data Dissemination.” The title of the article is “User-Centered Design and Innovation in the Sociometrics Social Science Electronic Data Library (SSEDL).” The authors, Josefin J. Card, Tamara Kuhn, and Thomas Wells, are all at Sociometrics Corporation. The article describes the Sociometrics Data Archives SSEDL as being a rich source of data for those in the public health, medical, nursing, social work, and social science professions. In the “product package,” datasets come with several data input files for SAS and SPSS. Purchasers can acquire data on CD-ROM or download from the web site, and downloading has had a significant rise in usage. In order to help users identify relevant datasets, Sociometrics has launched a topic-based drill down system showing “areas of richness,” which helps users identify and reach datasets and variables and the appropriate documentation.

The last article is “Overview of a Proposed Standard for the Scholarly Citation of Quantitative Data,” by Micah Altman and Gary King from Harvard University. This is an extended abstract summarizing a proposed standard for citation. This was presented at the IASSIST 2006 conference at the session “New Standards in Statistics and Data Citations.” The authors mention that, at a minimum, citations should include author(s), date of publication of the data set, and the data set title. These fields have been discussed before during the close to 50 years of documentation of data sets, and the fields are not so unambiguous now not to be discussed further. For instance, a great many people can be called “authors” in the production of a data set, and the same data set can have several other relevant dates attached besides “publishing date.” The authors recommend using the DDI elements, but their main purpose is to propose some novel fields that are directly linked to the use of modern technology. First of all, a Unique Global Identifier. The authors mention a naming resolution service and that brings to mind the technology of the Internet with name servers for looking up the correct IP address; however in this context, more than one copy of the dataset can exist at different physical locations. Secondly, technology is applied by adding a Universal Numeric Fingerprint. This should guarantee that the dataset has not been changed even though the data set might exist in different software. This should probably apply to the documentation as well.

The IASSIST is always open at its website, http://iassistdata.org, where you can look at conference information and visit the IASSIST blog (IASSIST Communiqué, http://iassistblog.org).

Articles for the IASSIST Quarterly are most welcome. Articles can be papers from IASSIST conferences, from other conferences, from local presentations, discussion input, etc. Contact the editor via e-mail: kbr@sam.sdu.dk.

Karsten Boye Rasmussen, March 2007
Introduction
In recent years, the number of official microdata sets accessible as Scientific Use Files has increased significantly in Germany. These microdata are of great interest to both economists and social scientists but are not, however, easy to work with.

Official microdata are surveyed to meet the data requirements of the German Federal Statistical Office. The data contain special classification types which need to be documented for the user. Users from the scientific community require more than superficial descriptions of a particular dataset; they also require detailed information pertaining to every variable in it.

Such an example of a German information system designed to fulfill the need of researchers, is the Microdata Information System (MISSY) presented below1. MISSY contains metadata or “data about data”, (Jacobs 2006) about the German Microcensus. MISSY piloted a project containing the descriptions for two census years 1995 and 1997.

The next section introduces the Microcensus and describes the functions and benefits of MISSY. The last section introduces the steps necessary to fully implement the system. Questions concerning general rules for documenting and presenting metadata derived from the experiences in the first phase of the project are also addressed.

The German Microcensus
The Microcensus is the biggest continuing survey in Germany. Conducted annually since 1957 by the Federal Statistical Office, it samples one percent of all German households or approximately 820,000 people. The main topics of the Microcensus are occupation and qualification, labor markets and household and family structures. Every four years the Microcensus contains additional questions about health or housing conditions, for example. The large sample size and the broad scope of topics make the Microcensus an invaluable data source for different scientific questions of varying complexity. For example, the Microcensus enables one to examine higher education among relatively small groups of immigrants, e.g. Italians or Greeks.

The Microcensus cannot be accessed by the scientific community in its entirety, but the Federal Statistical Office extracts a 70 per cent subset and provides it to researchers.

The Microcensus has attributes that are not commonly included in social sciences surveys: the classifications of professions (KldB – Klassifikation der Berufe) and economic sectors (WZ – Wirtschaftszweige) are used only in the official statistics and require some explanation to the researcher. Another characteristic of the Microcensus is its vast quantity of derived variables, the so-called “Bandsatzerweiterungen und Typisierungen”, whereby the latter are based on different concepts of families and living arrangements. The generation of these variables is not easy to comprehend; again they are only partially accessible to the scientific community. As a result, to work competently and efficiently with the Microcensus data, the researcher requires information exceeding what a superficial description of the dataset can provide. For this reason MISSY was developed.

MISSY
MISSY is a product of the German Microdata Lab (GML), formerly named the Department for Microdata at the ZUMA (Centre for Survey research and Methodology) in Mannheim. Since the 1980s, the GML’s focus has been on the Microcensus and as part of this focus it has offered an array of comprehensive services to support use of the Files. The GML, in collaboration with the Federal Statistical Office, ensures that the procedures necessary for anonymizing the data to protect confidentiality are in place. All Files are checked prior to being released to the scientific community and comprehensive documentation of the Files is created. As well, the GML provides support to researchers by offering advice on both the methodology and the content. To facilitate work with, e.g., classifications unique to the Microcensus, microdata tools are developed. Finally, and of equal importance, the GML organizes user conferences and workshops to promote the advantages of the Microcensus data for scientific research and to enable and increase the opportunity for scientific communication among researchers (Lüttinger et al. 2004).
MISSY facilitates research based on the German Microcensus. Gathering all the necessary metadata incorporating the knowledge of the GML is the first step; this includes official documents of the Federal Statistical Office. The second step is one that connects all the metadata in a way that considers the textual relationships between the data and the enquiries of social scientists and economists. The implementation accomplished by MISSY is based on the DDI (Data Documentation Initiative) 2.1 standard.

MISSY is an exclusively German system; there are two reasons for it being unilingual. At first, researchers are forbidden from using the data abroad. The second and more important reason is that all documents and descriptions of the Microcensus are in German making a knowledge of the German language essential. For demonstration purposes the most important expressions in the following examples have been translated.

To classify the metadata type, MISSY utilizes the categories of Sundgren’s dimensions. The metadata for the Microcensus includes both pragmatic and semantic as well as syntactic aspects (Fischer 2005, Sundgren 2003). This means that MISSY encompasses data answering questions of why (pragmatic aspects), what (semantic aspects) and how (syntactic aspects). In the Microcensus documentation, it is helpful to differentiate between general information about the entire study and specific information about the variables. The difference between the elements “study description” and “data description” is found as defined in accordance with the “Data Documentation Initiative” (Jacobs and Thomas 2006).

Note that in the middle of the screen there are multiple access points for retrieving specific information. Furthermore, there is a brief overview of the function of MISSY and there are also links to more information about the Microcensus and MISSY. In addition to this “main entrance” for access to specific information, the short list on the left sidebar of the screen under the red header “Variableninformationen” (specific information) can be used as well. The second list with the green header “Allgemeine Informationen” contains general information only about the Microcensus: an introduction, questionnaires, codebooks, interviewer guides, frequencies and some tips and recommendations on working with the data. Information about classifications used by the Federal Statistical Office or the scientific community is included here. For an easier navigation of the MISSY pages all specific information has red headers and all general information has green headers.

**Points of Access**

The different points of access were designed to simplify the search for variables while recognizing the varying needs and skills of the users. The first point of access is a list of all variables, subdivided by census year. This is useful when information about a specific variable for a particular year is required and it is preferable that the user already has some knowledge of the structure of the Microcensus.

An easier way of access, albeit longer, is given by the thematic structure illustrated below:

Thematic access is appropriate when the researcher is
interested in a specific subject or field of research and wants to know if the Microcensus has relevant content. To start with, the researcher may choose from eleven topics leading to the secondary level; at this point there are two links. The first is to publications based upon the Microcensus that pertain to specific subjects, “ethnic minorities and migration” being one example (see fig. II). This makes it easy for the investigator to determine what research might be undertaken or see what research has already been done based on the Microcensus. The second link connects to tables containing examples of analyses. Again, using “ethnic minorities and migration” as an example, the user will find multiple tables including one which shows a comparison of the graduation rates between the German and Turkish population. The tables were created to assist novice data users, e.g. students. The aim is to encourage researchers and future researchers to use official microdata in their analysis.

The fastest method for obtaining specific information is via a matrix containing all variables for every year covered by the Scientific Use Files (see fig. III). The variables names in the matrix cells are linked to specific information about the variables. Furthermore, the matrix provides an overview of characteristics surveyed in specific years.

Because the Microcensus is conducted annually, it can be used to address questions requiring a consistent long-term view to observe social change in society.

In order to examine longer time periods the researcher requires information about the comparability of the variables over a specific timeframe. The matrix presents the most important changes that have occurred in the variables.

There was a significant change to the Microcensus questionnaire in 1996; many variables were split into two, marked in the matrix. The changes are indicated and explained in the tool tips. When it is possible to generate comparisons of variables, links to SPSS-Syntax are provided. With these instructions, comparisons between many of the variables, both before and after 1996, can easily be made.

Specific information: variables documentation
For each variable, all available metadata are centralized on a single site. Not only are the variable labels included, but also the text of the questions and related notations, if available. There is information about the guiding filters of the questionnaire or what attributes the respondent had to fulfill in order to be asked this special question. The value labels and frequencies give first impressions of the variable’s distribution. In the first lines of the variable description are links to shortcuts to detailed information of comparable variables for other years and for different levels of the thematic structure. These links are marked with red buttons to create visual consistency with the list containing the different possibilities of access to specific information. Analogously, the links to general information that could be
of interest according to the particular variable are marked with green buttons. With these links, the researcher will be directed to the exact reference in the questionnaire, the codebook or the interviewer guide that contain the information concerning the variable of interest.

As stated in the introduction, the Microcensus contains a variety of derived variables which cannot be tracked in their composition. Information about the generation of these variables is documented for the internal use of the Federal Statistical Office only and cannot be accessed via the Internet. MISSY provides an additional link from the special information about derived variables that point to a site on which the generation of this particular variable is described. If the generation of the variable is based upon a special concept of families, households or living arrangements used by the Federal Statistical Office, another link to a description of these concepts is provided. Furthermore, researchers can go to a catalogue that contains definitions of the terms used in the Microcensus. Below is an example for a description of the variable “Type of working hours of the reference person of the family”:

This information makes it relatively easy to use even rather complicated variables in an appropriate way.

**Conclusion**

What conclusions can be drawn about the implementation of an information system for microdata? First of all, the concept of the system requires knowledge and research experience with the particular data. The special characteristics of the data should be understood and adequately documented. Secondly, knowing the data should make it possible to connect different kinds of information about the contents and thereby facilitate the search on particular topics. Ideally researchers should find not only all information they are looking for but also other helpful information that they may not even know existed.

Another important point to appreciate after the first
implementation of an information system of course is to ensure a maximum of usability. The next step is to have experts and users of Microcensus data analyze MISSY’s performance and make recommendations for improvement. Following this, they are plans to extend MISSY by including all available Microcensus Scientific Use Files. Two more specialized files will be added: the Panel File and the Regional File. Because of the concept of the Microcensus as a rotating panel the Panel File would include four years of census microdata. The Regional File contains microdata of a very differentiated regional level but on a less differentiated topic level to ensure the necessary confidentiality.

To include these new types of data sets in MISSY adequately, some modifications to the information system will be required. Another emphasis will be the extension of the category “tables” that provide an overview of the research possibilities using the Microcensus. An exercise-based introduction into working with Microcensus data is planned. With this concept the main focus of collecting and providing metadata for datasets will be expanded in a direction with implications for more practical and concrete advice for special problems that arise when working with Microcensus data. The result will be that the proportion of metadata with syntactic aspects will increase in MISSY.

References


* Andrea Janssen and Jeanette Bohr. Contact: Andrea Janssen, Centre for Survey Research and Methodology, ZUMA P.O. Box 12 21 55 - 68072 Mannheim - Germany. email janssen@zuma-mannheim.de.

Footnotes
1 http://www.gesis.org/Dauerbeobachtung/GML/MISSY/

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

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

<table>
<thead>
<tr>
<th>Topically-Focused Archive</th>
<th>Studies</th>
<th>Datasets</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adolescent Pregnancy</td>
<td>162</td>
<td>286</td>
<td>80,000</td>
</tr>
<tr>
<td>Aging</td>
<td>3</td>
<td>22</td>
<td>19,000</td>
</tr>
<tr>
<td>Child/Well-Being &amp; Poverty</td>
<td>12</td>
<td>36</td>
<td>20,000</td>
</tr>
<tr>
<td>Complementary &amp; Alternative Medicine</td>
<td>8</td>
<td>17</td>
<td>10,000</td>
</tr>
<tr>
<td>Contextual</td>
<td>13</td>
<td>29</td>
<td>19,000</td>
</tr>
<tr>
<td>Disability</td>
<td>19</td>
<td>40</td>
<td>25,000</td>
</tr>
<tr>
<td>Family</td>
<td>20</td>
<td>122</td>
<td>66,000</td>
</tr>
<tr>
<td>HIV / AIDS / STD</td>
<td>19</td>
<td>30</td>
<td>19,000</td>
</tr>
<tr>
<td>Maternal Drug Abuse</td>
<td>7</td>
<td>13</td>
<td>5,000</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>263</strong></td>
<td><strong>595</strong></td>
<td><strong>263,000</strong></td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th>Year</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visits to Sociometrics website</td>
<td>136,598</td>
<td>145,237</td>
<td>175,984</td>
<td>182,528</td>
<td>253,636</td>
</tr>
<tr>
<td>Hits to all data archive pages</td>
<td>42,003</td>
<td>70,022</td>
<td>86,770</td>
<td>101,819</td>
<td>210,494</td>
</tr>
<tr>
<td>Downloads of data and data-related products</td>
<td>10,704</td>
<td>16,325</td>
<td>29,729</td>
<td>34,954</td>
<td>39,672</td>
</tr>
<tr>
<td>Number of Units Ordered (non-subscribers)</td>
<td>171</td>
<td>222</td>
<td>156</td>
<td>167</td>
<td>193</td>
</tr>
<tr>
<td>Number of Purchasers (non-subscribers)</td>
<td>82</td>
<td>105</td>
<td>91</td>
<td>98</td>
<td>138</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Topic and Type Distribution Search Matrix Interface (partial view)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Complementary and Alternative Medicine Data Archive</strong></td>
</tr>
<tr>
<td><strong>Topic and Type Distribution</strong></td>
</tr>
<tr>
<td><strong>Mouseover any topic or type to view a definition.</strong></td>
</tr>
<tr>
<td><strong>Table</strong></td>
</tr>
<tr>
<td><strong>Column Headers</strong></td>
</tr>
<tr>
<td><strong>Behavior</strong></td>
</tr>
<tr>
<td><strong>Clinical Diagnosis</strong></td>
</tr>
<tr>
<td><strong>Cognition</strong></td>
</tr>
<tr>
<td><strong>Emotion</strong></td>
</tr>
<tr>
<td><strong>History</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Acupuncture</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Alternative Medical Systems, Other</td>
</tr>
<tr>
<td>Biological Function, Reproduction</td>
</tr>
<tr>
<td>CAM Therapies, General</td>
</tr>
<tr>
<td>Chiropractic</td>
</tr>
<tr>
<td>Conventional Therapies</td>
</tr>
<tr>
<td>Psychological Function, Development</td>
</tr>
<tr>
<td>Quality of Life, General Health</td>
</tr>
<tr>
<td>Race, Ethnicity</td>
</tr>
<tr>
<td>Reflexology</td>
</tr>
<tr>
<td>Region, State</td>
</tr>
<tr>
<td>Religion</td>
</tr>
<tr>
<td>Supplements, Vitamins</td>
</tr>
<tr>
<td>Traditional Chinese Medicine</td>
</tr>
<tr>
<td>Wealth, Finances</td>
</tr>
<tr>
<td><strong>Total</strong></td>
</tr>
</tbody>
</table>

Figure 1. Topic and Type Distribution Search Matrix Interface (partial 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 for


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: 1

0.80 CAM 17 Variable: MDB17097 Ever seen hypnotherapist


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
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.

---

**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)**:

<table>
<thead>
<tr>
<th>Value</th>
<th>Value Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>1</td>
<td>Yes</td>
</tr>
</tbody>
</table>

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 SISED 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 Josefin J. Card, Tamara Kuhn, and Thomas Wells are all at Sociometrics Corporation. Correspondence to: Dr. Josefi na J. Card, Sociometrics Corporation, 170 State Street, Suite 260, Los Altos, CA 94022, (650) 949-3282 x211, jjcard@socio.com.

**References**


Overview of a Proposed Standard for the Scholarly Citation of Quantitative Data

Critical components of the scholarly and library community are use of a common language and universal standards for scholarly citations and credit attribution, to enable the location and retrieval of articles and books. We present a proposal for a similar universal standard for citing quantitative data that retains the advantages of print citations, adds other components made possible by, and needed due to, the digital form and systematic nature of quantitative datasets, and is consistent with most existing subfield-specific approaches. Although the digital library field includes numerous creative ideas, we limit ourselves to only those elements that appear ready for easy practical use by scientists, journal editors, publishers, librarians, and archivists.

We propose that citations to numerical data include, at a minimum, six required components. The first three components are traditional, directly paralleling print documents. They include the author(s) of the data set, the date the data set was published or otherwise made public, and the data set title. These are meant to be formatted in the style of the article or book in which the citation appears. The author, date, and title are useful for quickly understanding the nature of the data being cited, and when searching for the data. However, these attributes alone do not unambiguously identify a particular data set, nor can they be used for reliable location, retrieval, or verification of the study. Thus, we add three components using modern technology, each of which is designed to persist even when the technology inevitably changes. They are also designed to take advantage of the digital form of quantitative data.

The fourth component is a unique global identifier, which is a short name or character string guaranteed to be unique among all such names, that permanently identifies the data set independent of its location. We allow for any naming scheme to be chosen, so long as it (1) unambiguously identifies the data set object, (2) is globally unique, and (3) is associated with a naming resolution service that takes the name as input and shows how to find one or more copies of the identical data set. Long-term persistence of the resolution service is meant to be guaranteed by the organization that operates it, although it is now becoming common to set up redundant multiple naming resolution services, so that archives can back each other up in case one goes out of business.

Unique global identifiers guarantee persistence of the link from the citation to the object, but we also need to guarantee and independently verify that the object does not change in any meaningful way, even when data storage formats change. To address this need, we add as the next component a Universal Numeric Fingerprint, or UNF. The UNF is a short, fixed-length string of numbers and characters that summarize all the content in the data set, such that a change in any part of the data would produce a completely different UNF. A UNF works by first translating the data into a canonical form with fixed degrees of numerical precision, and then applies a cryptographic hash function to produce the short string. The advantage of canonicalization is that UNFs (but not raw hash functions) are format-independent: they keep the same value even if the data set is moved between software programs, file storage systems, compression schemes, operating systems, or hardware platforms. Finally, since most web browsers do not currently recognize global unique identifiers directly (i.e., without typing them into a web form), we add as a final component of the citation standard a bridge service, which is designed to make this task easier in the medium term. Given how web services are accessed presently, the bridge service should be a URL, which can thus be recognized by any browser.

We also offer a systematic way to add information to data citations that also retains complete flexibility in added content. For each added element, we recommend a two-part syntax composed of: the value of the content, a field name that describes the content being added, and an (optional) semicolon separator. For example: “value [fieldname];” or “ Interuniversity Consortium for Political and Social Research [Distributor];”. To encourage standardization, we recommend that field names be drawn from the DDI 2.1 specification elements for study and variable descriptions. If others are needed, additional items may be drawn from other metadata schemes and vocabularies by adding the identifier for that scheme in parentheses within the bracketed field name, such as “Dataset [Type (DC)]” or “Current Population Survey Supplements [Series (ISO 690-2)]”. In unusual cases, users
could even easily add their own vocabulary if needed. This extended standard can be used to create citations similar to and compatible with some existing approaches, such as ISO 690-2 (see ISO, 1997) (although some aspects of these approaches may now be obsolete).

Together, the global unique identifier, UNF, and bridge service ensure permanence, verifiability, and accessibility even in situations where the data are confidential, restricted, or proprietary; the sponsoring organization changes names, moves, or goes out of business; or new citation standards evolve. Together with the author, title, and date, which are easier for humans and search engines to understand, all elements of the proposed full citation for quantitative data should achieve what print citations do and, in addition to being somewhat less redundant, take advantage of the special features of digital data to make the citation considerably more functional.

* This extended abstract summarizes the proposed standard. This was presented at the IASSIST 2006 conference in Ann Arbor at the session “New Standards in Statistics and Data Citations” by Micah Altman, Harvard University. Micah Altman is Associate Director, Harvard-MIT Data Center & Senior Research Scientist, Institute for Quantitative Social Science; Harvard University. Gary King is David Florence Professor of Government, Institute for Quantitative Social Science, Harvard University.

Our thanks to the National Science Foundation (SES-0318275, IIS-9874747) and the National Institutes of Aging (P01 AG17625-01) for research support.

The International Association for Social Science Information Service and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the IASSIST QUARTERLY. They also benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST.

Membership fees are:
- Regular Membership: $50.00 per calendar year.
- Student Membership: $25.00 per calendar year.

Institutional subscriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subscription: $75.00 per calendar year (includes one volume of the Quarterly).

Please make checks payable, in US funds, to IASSIST and Mail to:
IASSIST,
Assistant Treasurer US
Shelagh Mackay
University of California
Davis, CA, USA
95616-8617