ICPSR BULLETIN

Spring 2006 Vol. XXVI, No. 2

Felicia B. LeClere
Data Sharing for Demographic Research at ICPSR

Jeri Schneider
Why We Need a Data Citation Standard: Lessons Learned From Compiling ICPSR's Bibliography of Data-Related Literature

➢ Long-Time Summer Program Instructor, Jim Dowdy, Retires

➢ Register Now for IASSIST 2006

➢ New Council Members Elected

➢ Summer Training Program Offers New Courses

ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH
A PARTNER IN SOCIAL SCIENCE RESEARCH

Institute for Social Research • The University of Michigan
426 Thompson Street • Ann Arbor MI 48106-1248

ADMINISTRATION & STAFF

Myron Gutmann, Director
gutmann@umich.edu (734) 615-8400

Rita Bantom, Manager, Human Resources & Facilities
rybantom@umich.edu (734) 647-5000

Stacey Kubitz, Manager, Finance and Research Administration
skubitz@umich.edu (734) 615-2315

Bree Gunter, Membership Coordinator
bgunter@umich.edu (734) 615-8400

COLLECTION DEVELOPMENT

Erik Austin, Director
ewa@umich.edu (734) 615-7667

Peter Granda, Assistant Director
peterg@umich.edu (734) 615-2977

Peter Joftis, Director
Child Care and Early Education *Research Connections*
joftis@umich.edu (734) 615-3793

Felicia B. LeClere, Director
Data Sharing for Demographic Research
fleclere@umich.edu (734) 615-7333

JoAnne McFarland O'Rourke, Director
Substance Abuse and Mental Health Data Archive
jmcfar@umich.edu (734) 615-9528

James McNally, Director, Program on Aging
jmcnally@umich.edu (734) 615-9520

Christopher D. Maxwell, Director
National Archive of Criminal Justice Data
cmaxwell@umich.edu (517) 347-8081

Amy Pienta, Acquisitions Director
apienta@umich.edu (734) 615-7957

EDUCATIONAL RESOURCES

Henry Heitowit, Director
heitowit@*umich.edu* (734) 763-7400

Dieter Burrell, Assistant Director
dburrell@umich.edu (734) 763-7400

COMPUTER & NETWORK SERVICES

Bryan Beecher, Director
bryan@umich.edu (734) 615-2976

Asmat Noori, Assistant Director
asmat@umich.edu (734) 615-6386

DATA SECURITY & PRESERVATION

Cole Whiteman, Interim Director
colew@umich.edu (734) 615-7935

COLLECTION DELIVERY

Mary Vardigan, Director
vardigan@umich.edu (734) 615-7908

Mary Morris, Director, User Support
memorris@umich.edu (734) 647-2200

Linda Detterman, Director
Marketing & Promotion
lindamd@umich.edu (734) 615-5494

COUNCIL MEMBERS, 2006–2008

ICPSR BULLETIN

Spring 2006 • Volume XXVI, No. 2

The Inter-university Consortium for Political and Social Research (ICPSR), located at the Institute for Social Research at the University of Michigan in Ann Arbor, is the world's largest repository of computer-readable social science data. For over 40 years, the Consortium has served the social science community by acquiring, processing, and distributing data collections on a broad range of topics. Researchers at the Consortium's member institutions may obtain any of these data collections at no charge; researchers at nonmember institutions may also use the data, after paying an access fee. To find out more about ICPSR's holdings or about a specific data collection, visit the ICPSR Web site at www.icpsr.umich.edu.

The *ICPSR Bulletin* is published twice a year to inform interested scholars, including Official Representatives at the member campuses and ICPSR Council members, about topics and events relevant to ICPSR and its data collections. For subscription information, please contact the Editor.

Subscription rate: $15 per year

Editor: Ruth Shamraj
Assistant Editor: Naghma Husain

Data Sharing for Demographic Research at ICPSR

Felicia B. LeClere

ICPSR, University of Michigan

Among the most widely held scientific norms is that research should be replicable. Transparency in the research process allows other scientists to understand and reproduce both the methodology and the results of published studies. To that end, sharing data collected in the course of research is critical to the scientific endeavor, and many scholarly journals now require that authors submit relevant data files along with their manuscripts or make the data available for inspection upon request. In addition to being a sound practice scientifically, this process also serves as protection against misconduct.

With the growing emphasis on data sharing, expectations about data quality and availability have also changed dramatically. Data analysts expect to have quick access to secondary data that are well-documented and easily managed. These changing expectations, along with the fast pace of technological change, represent significant challenges to researchers who are collecting data with substantive goals and want to disseminate them for secondary use.

Response to Changing NIH Policy

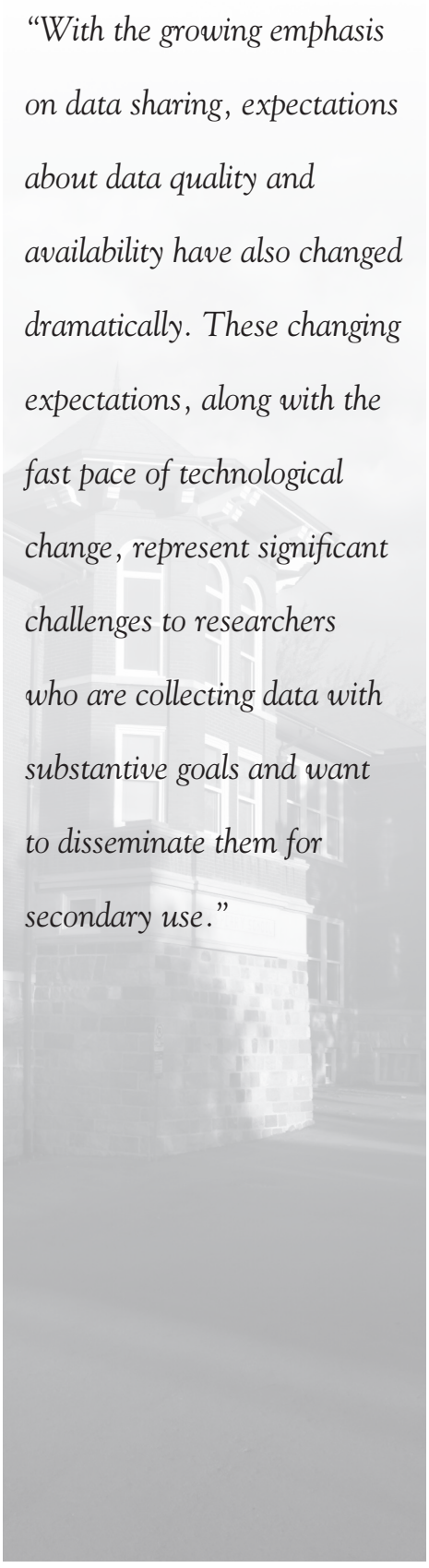
The federal government has recently made explicit its recommendations regarding data sharing in scientific research. On October 1, 2003, the National Institutes of Health (NIH) issued guidelines about sharing data collected during the course of federally funded projects.[1] Grant applications submitted on or after that date with a request for $500,000 or more in direct costs require a data sharing plan. This initiative is part of a larger focus at NIH and the National Science Foundation to increase the availability of scientific data to the scientific community.[2]

Responding to this emphasis on data sharing, ICPSR, along with partners at the population research centers of the University of Michigan, University of North Carolina, and University of Minnesota, submitted a proposal to NIH (Request for Application HD-03-032) entitled "Infrastructure for Data Sharing and Archiving." The call for applications recognized that sharing data collected under the auspices of research activity is often challenging and not well supported. Further, preparing data for release to a larger scientific community poses many challenges that researchers may have not faced before, including disclosure risk review, the preparation of documentation and user support material, restricted data contract development, and responding to direct user inquiries. The release of data is even more challenging when the data have a complex structure and when maintaining respondent confidentiality is particularly problematic. The application was successful, and NIH has funded a cooperative agreement called Data Sharing for Demographic Research (DSDR), which is now in its second year with leadership by Myron Gutmann (ICPSR), Steve Ruggles (Minnesota Population Center), Barbara Entwisle (Carolina Population Center) and project manager Felicia LeClere (ICPSR).

DSDR's cooperative agreement has four main aims. The first is to provide enhanced archiving capacity for projects

Felicia LeClere is the director of the ICPSR project Data Sharing for Demographic Research, and holds a cross-appointment as Associate Research Scientist in the Population Studies Center at the Institute for Social Research. Her research interests include the health of immigrants and the influence of social and physical geography on health disparities. Dr. LeClere was a team leader for the 1996 redesign of the National Health Interview Survey and headed the data center for a four-site, five-year longitudinal study of child abuse and neglect at the University of Notre Dame. She was also the Director of the Laboratory for Social Research, which provided social science data and statistical support to the University of Notre Dame.

"With the growing emphasis on data sharing, expectations about data quality and availability have also changed dramatically. These changing expectations, along with the fast pace of technological change, represent significant challenges to researchers who are collecting data with substantive goals and want to disseminate them for secondary use."



funded by the National Institute of Child Health and Human Development (NICHD), Demographic and Behavioral Sciences Branch (DBSB). The second aim is to provide data producers with a collection of tools to assist them in preparing complex data for release. The third aim is to improve access to complex demographic data by providing users with tools to expedite access and analysis. The final objective of the agreement involves research and development in areas related to improved data access and the protection of confidentiality.

The structure of the cooperative agreement calls for participation by the sponsoring agency, the partners in the agreement, and the research community at large. DSDR has an advisory group made up of members of the demographic research community and government agencies including the National Science Foundation, the National Institute on Aging, and NICHD. The role of the advisory panel is to help shape the priorities of DSDR and to monitor the implementation of major agreement objectives. The structure of the agreement also includes a technical advisory group that reviews progress and provides technical guidance to DSDR staff. This group consists of the project principal investigators and affiliates, DSDR staff, and Christine Bachrach, Branch Chief of DBSB, NICHD. Because this project is designed to serve the larger demographic community, DSDR principal investigator Myron Gutmann and project manager Felicia LeClere will, over the course of the project, visit all of the population research centers funded through the R24 mechanism by NICHD.

Data Archiving Options

In the traditional data archiving model, an archive like ICPSR holds and disseminates to the public the physical data and documentation files comprising any given data collection. This archiving method represents the bulk of ICPSR's archival holdings. However, for active data collection projects in which ongoing data management and dissemination are important components of the research activity, transferring the data to ICPSR may not be immediately appropriate. This means that the tasks of timely archiving, dissemination, and user support may fall to the data producers. DSDR, along with other units at ICPSR, is developing additional archival options for data producers in these situations. We provide producers with five primary alternatives to the traditional archive model:

- Data preservation only
- Data preservation with delayed dissemination
- Restricted-use data
- Enclave release of data
- "Virtual" data archiving

Public release of data is the most common ICPSR data release mechanism. The physical data files are available for download from ICPSR with enhanced documentation, setup files in SAS, SPSS, and Stata, ready-to-go files in these formats, and related literature citations where relevant. Archiving for **preservation only** implies that ICPSR will provide data producers with the option to store a copy of the final data file that is never distributed to the public. The preservation copy will undergo ICPSR media migration with changes in technology so that data producers can be assured of a usable electronic copy in the future. **Preservation with delayed dissemination** allows data producers to archive data while staff are still active on the project, but provides a protected time for research activity to be completed. A release schedule is negotiated with producers

at the time of deposit to schedule the public dissemination of the data file. The **restricted-use** and **enclave release** alternatives allow producers to turn over to DSDR the responsibility of designing and monitoring the release of confidential data through contractual arrangements. Data producers may wish to retain control of the public-use data files but forego the dissemination of confidential files, which require both careful control and constant monitoring. ICPSR offers carefully structured restricted-use contracts as well as the option to issue data through a secure data enclave in Ann Arbor. Finally, the **virtual archiving** option offers producers a selection of virtual options currently in use at ICPSR that can improve the visibility and usability of producer-disseminated data to users. The package of services includes union catalog listing, full-text linked bibliography, user registration and monitoring, and linked Web access. Thus, a data producer can retain control of data dissemination but DSDR will provide access through the ICPSR search interface. To date, DSDR has created union catalog entries and compiled related literature bibliographies for 15 studies, the majority of which are held elsewhere. Transitional Web pages are being developed that will standardize the linkage between ICPSR and the producers' data dissemination Web pages.

These flexible archiving options allow data producers to pursue archival strategies that fit both their funding mechanism and planned research activity. Many funded data collection projects and primary investigators do not have the resources or infrastructure necessary to disseminate data to a large user public and may decide to select ICPSR as the primary source for dissemination. Alternatively, projects that incorporate data dissemination into planned research activity may choose to use ICPSR's virtual archiving services to increase the visibility of new data files. DSDR assists data producers in making the most appropriate choice for their projects.

In addition to expanding archiving options, DSDR is also participating in an effort to identify and track data collected under the auspices of federal grants. As an outgrowth of another ICPSR project using information from the NIH Computer Retrieval of Scientific Information (CRISP) database, DSDR is screening all funded grant abstracts from NICHD for the years 1972 through 2005.[3] After an initial screening step used to identify funded social and behavioral science research, project staff at ICPSR are reviewing abstracts for those projects of specific interest to demographers in which data were collected. We have developed a standardized protocol to identify projects of topical interest. Once we have completed the screening, we will attempt to locate data files that may have resulted from the grant activity. In cases where the data files are not available in the public domain, we will contact the principal investigators to inquire about archiving opportunities. We are also reviewing P01 and R01 abstracts from 1979–2005 from the Demographic Behavioral Sciences Branch of NICHD in the same manner to identify primary and secondary data collections. This information will help set our archiving priorities.

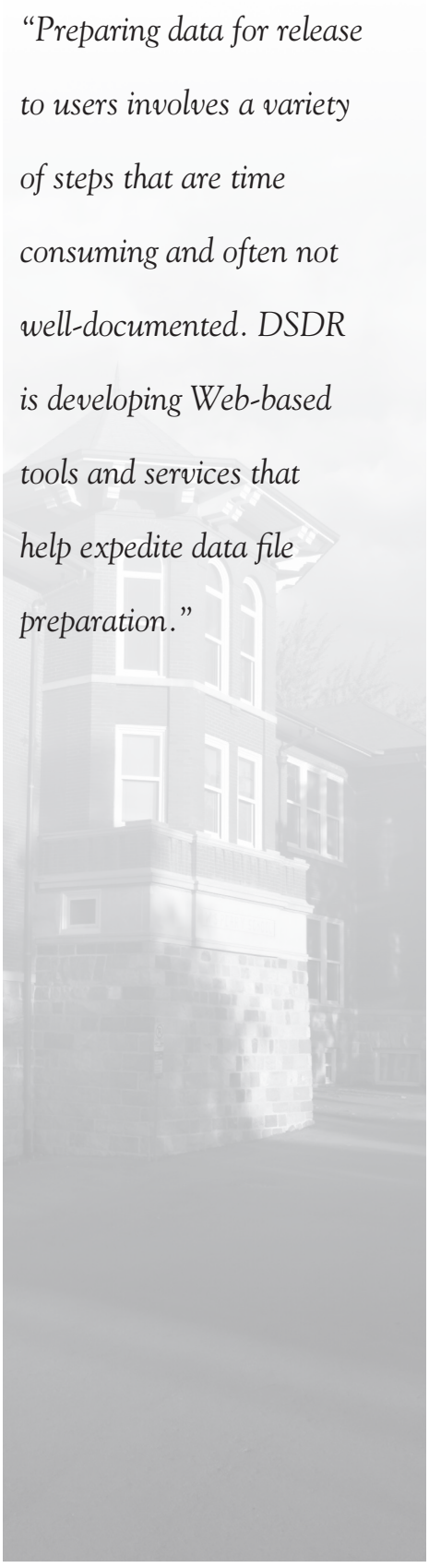
Tools for Producers

Preparing data for release to users involves a variety of steps that are time consuming and often not well-documented. This is especially true when the data collection activities and resulting data files are complex. DSDR is developing Web-based tools and services that help expedite data file preparation. We currently have four such projects in various stages of development.

User guide preparation. One of the most difficult and often neglected steps in the preparation of data for release is the preparation of user guides. Data producers often feel that survey design documents, variable and file documentation, and the data file itself should contain enough information for new data users to conduct analyses. Experience with complex data files suggests that this is in fact not the case. Large data systems such as the Survey of Income and Program Participation, the Health and Retirement Survey, and the National Longitudinal Study of Youth have invested substantial resources in developing a collection of guides that allow users to navigate complex file structures, longitudinal panels, and changing variable and design definitions. In a project with the Population Studies Center at the University of Michigan, DSDR is developing a Web-driven tutorial providing instruction on how to develop user guides. This tutorial will allow data producers to identify and define the potentially relevant topic areas as well as find helpful model guides and resources elsewhere on the Internet. Sections of the tutorial will also be useful as a structured outline for a standard user guide. The Web tutorial is planned for release in the summer of 2006.

Anonymizing qualitative data. Little technological support exists for researchers who wish to share qualitative data in a responsible manner with others. One of the primary tasks necessary to prepare qualitative data files for release is the process of "anonymization," or the systematic removal of identifying information from field notes and interviews. Retaining the consistency of names, relationships, and locations is a particular challenge.

"Preparing data for release to users involves a variety of steps that are time consuming and often not well-documented. DSDR is developing Web-based tools and services that help expedite data file preparation."



DSDR, in collaboration with Computer and Network Services at ICPSR, the University of Pennsylvania, and the Institute for Research on Women and Gender at the University of Michigan, will be developing customized "anonymization" software. Qualitative data from an actual research project will be used to develop and test the software. DSDR will subsequently make the data available. The software will then be used throughout ICPSR to assist with the "anonymization" of qualitative data. Work on software development began in January 2006.

Disclosure risk review. While the administrative review of data files for identity disclosure risk is a standard part of the process federal agencies in the United States use to prepare data for public release, this step in the process is less likely to occur in a standard way among research projects outside the federal system. Increasing public concern about the privacy of survey respondents suggests that data producers should be more routinely cautious about these issues. To assist data producers, DSDR has developed two related services to improve the review and monitoring of confidential data. The first service offered to data producers is a systematic confidentiality review of data files prepared for release. Using the Checklist on Disclosure Potential of Proposed Data Releases prepared by the Interagency Confidentiality and Data Access Group of the Federal Committee on Statistical Methodology,[4] as well additional revisions by a working group at ICPSR,[5] we conduct a review of existing data items and assess the potential risk for disclosure of identities. The review consists of a systematic evaluation of the documentation and data to assess the likelihood of attribute disclosure, which would entail finding individuals known to be in the survey through their characteristics, and an assessment of the potential for direct disclosure through matching external data files to survey data. We have completed our first such review for the Fragile Families and Child Wellbeing Study[6] and are currently conducting an analysis for the Welfare, Children and Families: A Three Cities Study.[7] These disclosure reviews are not intended to be prescriptive or definitive but are simply systematic guides for data producers as they develop their data for release or as they reassess the content of previous releases.

Restricted data agreements. If data producers deem that some data pose a risk for identity disclosure yet are of substantial analytical interest, they may choose to distribute the data through a restricted-use agreement. Restricted-use agreements have proliferated in recent years, yet little guidance is available to those interested in writing such agreements. DSDR has developed Web-based procedural and content guidelines to assist producers in this task. These guidelines address not only contract areas but also the degree of use restriction based on the sensitivity of the data. Data producers may craft overly restrictive agreements if they do not have enough information to weigh contractual requirements against the disclosure risk posed by the sensitive data items. The information provided in these DSDR guidelines will help producers make informed decisions about these issues before they submit their draft contracts to university or organizational officers for review. The work on these guidelines is complete and awaits review by the University of Michigan, Office of the General Counsel. As noted earlier, when data producers feel they do not wish take on the task of writing and monitoring their own restricted data use agreements, DSDR can do so for them.

Tools and Services for Data Users

Increasingly complex data collection objectives and methods have made data analysis more difficult despite substantial improvements in electronic access and computing power. Expediting the release of data to the research community is only the first step in the process of sharing data. DSDR is currently pursuing two primary methods of improving users' access to data.

Training. DSDR sponsors ICPSR Summer Program short courses on subjects of interest to data users. In 2005, we cosponsored a one-week session with the Child Care and Early Education *Research Connections* project at ICPSR entitled "Data Sharing and Dissemination: Making Your Data a Resource for Others to Use." This session was aimed primarily at data librarians and archivists who are charged with making data available and usable. In the summer of 2006, we will be sponsoring a one-week session at the University of North Carolina on the Nang Rong data system,[8] which includes information from 20 years of data collected in villages in Northeast Thailand. Complex data systems such as Nang Rong require substantial training for users who are not involved in the project to feel prepared to undertake analyses. In subsequent years, we hope to identify additional data systems of this type and provide similar training.

Tutorials. We will also be developing user-based tools similar to our tools for data producers. The first project in the planning stage is to build tutorials on especially complex data access issues such as linking identifiers in longitudinal data files, harmonizing variable changes in repeated time series and longitudinal data files, and creating event files from individual survey records. The second project, also in the planning stage, is to provide short user guides for widely-used comparative demographic data. These guides will help users quickly assess the scope and sample of similar data files.

Enhanced documentation. We are also currently working with the staff of the Michigan Census Research Data Center and the Center for Economic Studies and Population Division of the U.S. Census Bureau to create electronic documentation and setup files for the 1960–2000 restricted decennial census microdata files. These data files, which are available upon application to the Census Research Data Centers, lack full and accessible documentation. We will be working with the Census Bureau to make enhanced documentation available through the ICPSR Web site. Staff of the Population Division at the Census Bureau began work on this project several years ago and DSDR staff will be assisting with its completion.

Research and Development

As part of the cooperative agreement that spawned DSDR, the project partners are pursuing research on technical and analytic projects related to data access, confidentiality, and documentation. The Minnesota Population Center is charged with developing a Web-based tool to improve access to complex data systems. The first step was to identify the most complex tasks associated with data preparation and analysis. They used both a survey of users and a targeted literature search to identify the demographic datasets and data preparation tasks users found the most difficult. They focused on six datasets including the National Longitudinal Study of Adolescent Health, the Fragile Families and Child Wellbeing Study, National Survey of Families and Households, Los Angeles Family and Neighborhood Survey (LAFANS), the Three Cities Studies (Welfare, Children, and Families: A Three-City Study Boston, Chicago, San Antonio) and the pilot for the New Immigrant Study. They identified four trouble spots, some of which they hope to target for an improved data access system. These problem areas included: (1) difficulty navigating the documentation; (2) reformatting and merging data files consistently; (3) recoding and constructing analytic variables; and (4) using the full scope (longitudinal, relational, contextual information) of the data files simultaneously. In the next year, the team will target at least one of these areas for development.

A substantial amount is known about disclosure risk associated with microdata and, as noted, a variety of methods are available to remedy these identified risks in the release of data. Less attention has been paid to the visual display of data from these same survey sources. Our Carolina Population Center collaborators have examined disclosure risk associated with the spatial display of data from sample surveys. In examples from the Longitudinal Study of Adolescent Health, a school-based survey, Barbara Entwisle and colleagues have demonstrated that the scale of spatial display and the heterogeneity of the spatial distribution of the sample cluster can substantially influence the likelihood of identifying schools. Their research describes the size of the spatial buffer necessary to avoid disclosure of a single sample unit.[9]

In partnership with the Data Documentation Initiative (DDI), an international alliance developing an XML standard for technical documentation, DSDR is helping to develop metadata specifications for longitudinal files. In a proposal recently submitted to the DDI Alliance and to partners in the International Association

"The staff of DSDR, and of ICPSR more generally, are committed to helping achieve the goal of scientific transparency by assisting data producers in preparing well-documented data and protecting the privacy of study participants."

for Social Science Information Service and Technology (IASSIST) working group on metadata standards, DSDR has made recommendations regarding the metadata elements necessary at the variable level to track changes in questions over time and those that are needed at the file level to track attrition of subjects over time. Staff at ICPSR and the Population Studies Center at the University of Michigan used an application of these metadata standards in a joint proposal currently under review to create an integrated time series for 11 fertility surveys in the United States beginning in 1955. These new metadata standards can be used to help assess the analytic implications of a changing sample of universes across surveys.

Conclusion

In summary, the staff of DSDR, and of ICPSR more generally, are committed to helping achieve the goal of scientific transparency by assisting data producers in preparing well-documented data and protecting the privacy of study participants. Creating flexible tools that allow producers who are actively collecting data to improve documentation and data quality and providing information and assistance to users to promote the use of extant data are integral objectives.

For more information about the Data Sharing for Demographic Research project, please contact us.

Web site: www.icpsr.umich.edu/DSDR
E-mail: dsdr@umich.edu
Phone: 734.615.4979

Project Manager:
Felicia B. LeClere, Ph.D.
734.615.7333 (phone)
734.647.8700 (fax)
fleclere@umich.edu

References

[1] National Institutes of Health. 2003. "Final NIH Statement on Sharing Research Data." Bethesda, MD: National Institutes of Health. Available: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

[2] Office of Extramural Research. 2003. "NIH Data Sharing Policy and Implementation Guidance." Bethesda, MD: National Institutes of Health. Available: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

[3] Computer Retrieval of Information on Scientific Projects (CRISP) (database). Bethesda, MD: National Institutes of Health. Available: http://crisp.cit.nih.gov/

[4] Office of Management and Budget. 2003. "Confidentiality and Data Access Committee (CDAC)." Washington, DC: Federal Committee on Statistical Methodology. Available: http://www.fcsm.gov/committees/cdac/

[5] Human Subject Protection and Disclosure Risk Analysis. (Web site). Ann Arbor, MI: Inter-university Consortium for Political and Social Research. Available: http://www.icpsr.umich.edu/HSP/

[6] The Fragile Families and Child Wellbeing Study. (Web site). Princeton, NJ: Center for Research on Child Wellbeing. Available: http://crcw.princeton.edu/ff.asp

[7] Welfare, Children, & Families: A Three City Study. (Web site). Baltimore, MD: Johns Hopkins University. Available: http://www.jhu.edu/~welfare/

[8] Nang Rong Projects. (Web site). Chapel Hill, NC: Carolina Population Center. Available: http://www.cpc.unc.edu/projects/nangrong/

[9] VanWey, Leah K., Ronald R. Rindfuss, Myron P. Gutmann, Barbara Entwisle, and Deborah Balk. 2005. "Confidentiality and Spatially Explicit Data: Concerns and Challenges." *Proceedings of the National Academy of Sciences* 102(43):15337–15342.

Why We Need a Data Citation Standard:

Lessons Learned From Compiling ICPSR's Bibliography of Data-Related Literature

Jeri Schneider
ICPSR, University of Michigan

After more than five years of work on the online ICPSR Bibliography of Data-Related Literature, we have accomplished and surpassed most of our original goals. The Bibliography currently contains over 38,000 citations to publications based on data in the ICPSR holdings. Two-way linkages enable users to identify a data collection of interest and link to its related literature, or to locate a publication and link to the underlying data used in analysis. Through OpenURL technology, the full text of many publications are delivered straight to the user's desktop, enhancing the Bibliography's utility.

In the process of constructing the Bibliography, we have also arrived at a new vision for its future, one that incorporates a data citation standard. What follows is a brief account of how we approached the Bibliography project and a description of the challenges we faced in compiling the resource. I conclude with a vision of how we may overcome these challenges and what our vision means for the future of data citation collection and analysis.

Bibliography Project History

The Bibliography, originally funded through a National Science Foundation "Infrastructure in the Social Sciences" award, first saw the light of day in late 2000. The newly compiled collection of citations drew from three main sources: (1) a dozen smaller, incomplete bibliographies culled from about 20 seminal social and political science journals and published in ICPSR Annual Reports from 1974 to 1985; (2) up to three "related publications" citations embedded in the individual study descriptions of hundreds of ICPSR datasets; and (3) a small collection of printed publications submitted by authors who had used ICPSR data.

The Bibliography project aimed to integrate these resources into a single collection and add to it the thousands of as-yet unharvested citations to publications based on data distributed by ICPSR since its inception in 1962. Project staff envisioned an online resource that would provide individual bibliographies for each dataset and be searchable in its entirety. We also wanted to provide a means and an incentive for authors to easily submit citations of their work based on ICPSR data. It was hoped that this comprehensive bibliography would serve as a tool for researchers, students, funding agencies, and other data users to better understand datasets and their contributions to the social sciences, and also track the uses and impact of data in ICPSR's holdings.

The first step in this endeavor was to select an application that could store and organize citations. We decided to

Jeri Schneider holds a Master of Science in Information degree from the University of Michigan, School of Information. She has worked at ICPSR as a librarian/ information specialist since 2000, managing the development of the Bibliography of Data-Related Literature and teaming with Web developers on usability issues. Her specialty is online search and retrieval.

"There is a growing movement among information specialists, researchers, data librarians, archivists, and other interested parties to establish a standard for citing data used in published works."

start with off-the-shelf bibliographic management software that would allow us to start collecting and organizing citations immediately, and then later migrate the citations to Oracle tables linked to the ICPSR Web site. We established a two-pronged approach to finding references to ICPSR datasets: we searched online resources for specific study titles and also identified a list of core journals, print and electronic, to browse. To date, we have covered the following ground in our research:

- Searched over 40 abstract/indexing and full-text databases for publications based on ICPSR's study titles
- Browsed the entire contents of over 60 journal titles published from 1962 to the present and browsed at least 100 other journals for a number of years
- Entered over 200 citations submitted by authors
- Browsed dozens of free Internet resources for working papers and agency reports
- Collected citations from a variety of other sources, including existing bibliographies

(For a full methodology report and a list of individuals involved in the Bibliography project, see "About the Bibliography of Data-Related Literature" at www.icpsr.umich.edu/citations/methodology.html).

The ICPSR Bibliography citations are stored in an Oracle database that is accessible through multiple entry points on the ICPSR Web site — through the searchable database itself, through the bibliographies for individual datasets and series studies linked from study descriptions, and through the bibliography files that users automatically download with each dataset.

In 2005, we enhanced the Bibliography by adding OpenURL links enabling users to link directly to their local library resources, including the full-text of journal articles. Staff members continue to identify and add new citations by browsing academic journals and searching online full-text and abstract/index databases for references to ICPSR study titles. A handful of authors also help populate the Bibliography by submitting about 25 citations per year for papers that use ICPSR data. All of these tasks are labor-intensive and require constant vigilance to keep pace with the prodigious output from academic publishing, conference presentations, and other sources.

Lack of a Data Citation Standard

The primary challenge to finding references to ICPSR data, aside from wading through the sheer volume of publications available, is the lack of a standard format for citing uses of data. Datasets are often not cited in papers at all, and when they are, the citations rarely appear in the references section. Instead, they haphazardly appear in the abstract, the methodology section, a footnote, or perhaps even in a caption to a graph or chart. When data are cited, because there is no standard format, the information provided often fails to accurately identify which data were used. To make matters even more difficult, different versions of a data collection may have been released over time. This means that ICPSR staff can spend countless hours scanning various sections of publications trying to determine whether a dataset was in fact used and then, if the indications are in the affirmative, identifying which dataset was used.

It should be common practice to cite data in a standard format within the references section of a paper. There are many reasons for following such a standard; the two most obvious are:

(1) to properly credit the producers of the data, and (2) to enable researchers to locate a dataset in order to replicate an analysis or perform different analyses on the same data. The publishing world long ago developed standards and mechanisms for properly citing references to publications, which has made it possible to develop "citation indexes" that are used to trace the impact of authors' papers and, perhaps more significantly, to enable the "association of ideas" among scholarly publications (see Garfield, 1955, for broader discussion of this topic in regard to science citation indexing). In the absence of a standard for citing data use, the social science research community lacks important tools for verifying and building upon prior research through replication and secondary analysis. The implementation of a standard data citation format will also enable the development of "data citation indexes," similar to print citation indexes, that would permit new analyses and associations to be made among datasets and publications based on those data.

Sue Dodd addressed the lack of data citation standards over 25 years ago in a seminal paper (Dodd 1979). She proposed a method for citing data and hoped that the guidelines would be included in such works as *The Chicago Manual of Style* and Turabian's *A Manual for Writers*. She also foresaw the inclusion of data citations in the Social Science Citations Index. A decade later, Dodd (1990) outlined a series of steps necessary to implement such a standard. However, little progress has been made to realize her vision of a citation standard and the benefits that would result.

Envisioning the Future: A Data Citation Standard and Beyond

Fortunately, there is a growing movement among information specialists, researchers, data librarians, archivists, and other interested parties to establish a standard for citing data used in published works. ICPSR recently developed a new data citation format that is included as part of each study description. We encourage authors to use this citation within published works; however, there is no mechanism to enforce the citation of our data. The International Association for Social Science Information Service and Technology (IASSIST) has established a Data Documentation and Citation Interest Group and initiated a discussion to facilitate agreement on a universal standard for citing data (Baxter 2005). Researchers at the Virtual Data Center, founded by the Harvard-MIT Data Center, are currently developing a standard method to create direct links from data citations in digital publications to online study descriptions available from data distributors. (See Altman and King 2005 for a description of the proposed standard.)

The key is to bring all of these stakeholders together with publishers, database vendors, and other interested groups to agree on a standard that will serve the various constituencies' needs and gain endorsements from the parties that are positioned to enforce the standard. Given the growing call for such a standard and the availability of digital resources to facilitate the creation and organization of citations, the time seems ripe to form a working group to carry out Dodd's suggestions for developing a standard for citing data. Dodd recognized the need to raise awareness among researchers and editors, to encourage publishers of "style manuals" to include guidelines for citing data, and to have social science information advocates involved in national and international standards committees to ensure that the needs of social science data users are included in discussions about citation standards.

Having worked for over five years developing ICPSR's Bibliography, we can see enormous potential in implementing a data citation standard. We would like to push the standard beyond the obvious purposes of crediting data producers and linking readers of articles to data. Ultimately, we see the need for the creation of an entirely new data citation indexing service that collects citations of publications that cite data, organizes them in a database, and provides links among a variety of online resources. This could

A data citation standard and subsequent development of an automated "data citation index" will make it possible to:

- Link quickly from electronic publications to data and vice versa
- Create dynamic bibliographies of individual datasets, series, or other categories of data collections
- Facilitate associations of ideas generated through the use of data
- Identify cross-disciplinary uses and implications of data analysis
- Track data usage as justification for continued funding of existing studies
- Demonstrate the need to fund new studies
- Demonstrate the impact of data collections
- Harness the potential of existing keyword/subject indexing of journal articles, as a means to search for relevant data
- Develop and explore new connections among uses of data

"We hope to participate actively in the movement to establish an international data citation standard and encourage the community to join us in this effort."

be an automated system that would "crawl" digital articles searching for data citations. When the crawler identifies a data citation, it would "harvest" that citation along with the citation for the publication in which it appears and pull that information into a "data citations index" database that could be used for a myriad of purposes (see sidebar). In order for such a system to work, data citations would not only have to be standardized, but also be in a machine-readable format and be readily identifiable as data citations, i.e., citations to datasets as opposed to publications.

Before such a system can be developed, however, agreement has to be reached as to what a data citation should look like — what elements must be present and what format should be used. A citation standard has to meet the varying needs of many parties, including principal investigators, data archivists and distributors, researchers, authors, publishers, students, funding agencies, and existing database vendors. Each of these groups may have different needs with regard to the citation. For example, principal investigators might want to use citations to track how their data have been used and how often. Students new to a dataset might want to see what has been done with the data so far, perhaps within a particular subject area. Researchers might use a citation to access the data for replication purposes. With direct links to underlying data, they can be assured that they are working with the intended version of the data. Funding agencies might be interested in the impact and uses of the data. Others might be interested in exploring cross-disciplinary uses of a dataset by tracking citations of the data.

Ironically, a data citation indexing service would eventually make ICPSR's Bibliography obsolete, at least in its present form. However, it would enable a much more robust and comprehensive collection of citations. Once a data citation standard is developed and implemented, and as more resources are made available electronically, manually searching for data citations will become less and less common.

In the meantime, ICPSR will continue to augment the Bibliography by adding citations and by developing new methods to serve the needs of data users. We hope to participate actively in the movement to establish an international data citation standard and encourage the community to join us in this effort.

References

Altman, Micah, and Gary King. 2006. "A Proposed Standard for the Scholarly Citation of Quantitative Data." Cambridge, MA: Harvard University. Available: gking.harvard.edu/files/cite.pdf

Baxter, Pam. 2005. "IASSIST Committees, Action Groups, & Interest Groups." Davis, CA: International Association for Social Science Information Service and Technology. Available: www.iassistdata.org/membership/committees.html

Dodd, Sue A. 1979. "Bibliographic references for numeric social science data files: Suggested guidelines." *Journal of the American Society for Information Science*, 30, 77–82.

Dodd, Sue A. 1990. "Bibliographic References for Computer Files in the Social Sciences: A Discussion Paper." Chapel Hill, NC: Institue for Research in Social Science. Available: www.people.virginia.edu/~pm9k/info/compRef.html

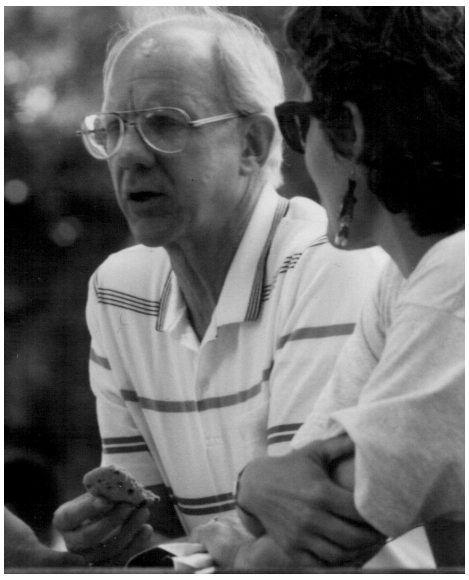
Garfield, Eugene. 1955. "Citation indexes for science: A new dimension in documentation through association of ideas." *Science*, 122, 108–111.

ANNOUNCEMENTS

Long-Time Summer Program Instructor, James Dowdy, Retires

Henry Heitowit
ICPSR, University of Michigan

After 26 continuous summers (1980–2005), James (Jim) Dowdy has retired from teaching in the ICPSR Summer Program. Jim was first hired as an instructor in the Program in 1980 when he was at West Virginia University. In 1994 he moved to the Department of Mathematics at Saint Louis University. Over these many years, Jim has excelled as the instructor for the summer course on "Mathematics for Social Scientists." In this lecture series, Jim unraveled the mysteries, utilities, and mechanics of matrix algebra. Jim's engaging teaching helped a generation of students gain competence in mathematics and statistics. He was a gifted and inspirational teacher. Many students commented that if they only had had such a wonderful math teacher in high school or college, their lives would have been different, and better for it. Across two-and-a-half decades, Jim taught well over 3,000 Summer Program participants. Jim was also a marvelous colleague and was greatly respected by his fellow Program instructors. Several years ago, he was unofficially dubbed the "Dean" of the Program faculty. He has been instrumental in creating, shaping, and evolving the Program curriculum and instructional environment. Partially in recognition of his seminal role in the Program, in 1993 he was presented the ICPSR Warren E. Miller Award for Meritorious Service to the Social Sciences. ICPSR is grateful for Jim's contributions to the Program. His dedication and talents have greatly enriched the Program and the lives of his students and colleagues.

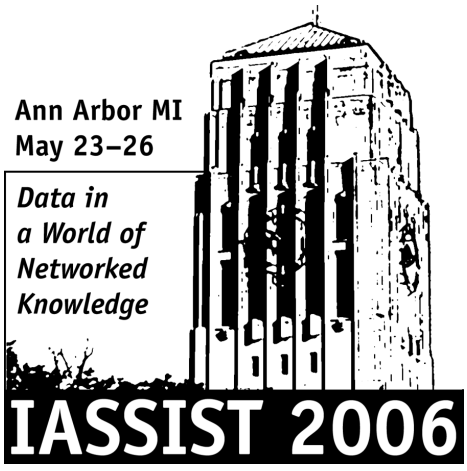


Retired Summer Program Instructor, James Dowdy

Register Now for IASSIST 2006

Online registration for the IASSIST 2006 conference, "Data in a World of Networked Knowledge," is now open. The conference will take place in Ann Arbor, Michigan, on the University of Michigan campus, on May 23–26, 2006.

This will be the 32nd annual conference of the International Association for Social Science Information Service and Technology (IASSIST). Participants will present papers, poster/demonstration sessions, and panel sessions on topics that address the full range of digital data life cycle issues, including access, documentation, dissemination, preservation, data use, and current empirical research activity. Additional topics will include information and statistical literacy, statistical disclosure, and Geographic Information Systems (GIS) and spatial data, as well as publication, annotation, curation, and authentication of networked knowledge assets.



The conference will be preceded by workshops on Tuesday, May 23, and followed by optional weekend activities in the Ann Arbor area. Details about the upcoming conference may be found on the IASSIST 2006 Web site: www.icpsr.umich.edu/iassist.

New Council Members Elected

ICPSR is pleased to announce the election by its membership of a new slate of ICPSR Council members, who will serve four-year terms beginning in March 2006. The election, which took place November 1–December 31, 2005, was conducted electronically, with a 34.3 percent participation rate.

The new members of Council include:

Rodolfo de la Garza, Political Science, Columbia University
Michael R. Haines, Economics, Colgate University
Kathleen Mullan Harris, Sociology, University of North Carolina, Chapel Hill
Aletha C. Huston, Human Development and Family Sciences, University of Texas
Nancy Y. McGovern, Digital Preservation, Cornell University
Samuel L. Myers Jr., Hubert Humphrey School, University of Minnesota

Ruth Peterson, Ohio State University, was elected Chair of the 2006–2008 Council. Mark Hayward, University of Texas, will serve as Past Chair for one year.

Continuing Council members, who will serve for two more years, include Darren W. Davis, Michigan State University; Charles H. Franklin, University of Wisconsin; Paula Lackie, Carleton College; James Oberly, University of Wisconsin, Eau Claire; and Walter Piovesan, Simon Fraser University.

Outgoing Council members are Ilona Einowski, University of California, Berkeley; John Handy, Morehouse College; and Ronald Rindfuss, University of North Carolina, Chapel Hill. Note that Nancy McGovern and Samuel Myers, who were completing terms of others on the previous Council cycle, have agreed to serve an additional term and were elected.

ICPSR thanks the outgoing Council members for their years of service, and warmly welcomes the incoming members.

Rodolfo de la Garza is Professor of Political Science at Columbia University. He also leads the Project on Immigration, Ethnicity, and Race, which is a part of Columbia's Institute of Social and Economic Research and Policy. In addition to his appointment in the Department of Political Science, de la Garza continues his work for the Tomas Rivera Policy Institute, a think tank housed at California's Claremont Graduate University that conducts policy research on issues affecting Latino communities. De la Garza combines interests in political behavior and public policy and is an expert on Latino political behavior and immigration. He has edited, coedited and coauthored numerous books including: *Sending Money Home: Hispanic Remittances and Community Development*; *Latinos and U.S. Foreign Policy: Lobbying for the Homeland?*; and *Bridging the Border: Transforming Mexico-U.S. Relations*. De la Garza served as a Vice President of the American Political Science Association and received the Life-Time Achievement Award of the Committee on the Status of Latinos in the Profession of the American Political Science Association in 1993. He has also served on the Executive Council of the Western Political Science Association. He was a founding member of the National Association of Chicano Studies and the Inter-University Program for Latino Research; he also directed the Inter-University Program for Latino Research/Social Science Research Council Joint Committee on Hispanic Policy Research; and he is a member of the Council of Foreign Relations.

Michael R. Haines is Banfi Vintners Professor of Economics at Colgate University and Research Associate at the National Bureau of Economic Research. He is an economic historian and historical demographer who works on historical fertility, mortality, and health in the United States and Europe, historical consumer behavior in the United States and Europe, and historical census and vital statistics materials. He has served as the treasurer, vice president, and president of the Social Science History Association; has served on several boards of editors; and has been a reviewer and consultant for NSF, NICHD, and the World Bank, and has been the recipient of several grants from NIH. He is the author of *Economic-Demographic Interrelations in Developing Agricultural Regions: A Case Study of Prussian Upper Silesia, 1840–1914*; *Fertility and Occupation: Population Patterns in Industrialization*; and *Fatal Years: Child Mortality in Late Nineteenth-Century America* (with Samuel H. Preston) and is coeditor (with Richard H. Steckel) and contributor to *A Population History of North America*. He is one of the editors-in-chief of the *Millennial Edition of the Historical Statistics of the United States*. He contributed the chapters on population and vital statistics to that edition.

Kathleen Mullan Harris is Gillian T. Cell Distinguished Professor of Sociology at the University of North Carolina at Chapel Hill and Faculty Fellow at the Carolina Population Center. Her research interests are in the areas of family, poverty, and social policy. Dr. Harris is Director of the National Longitudinal Study of Adolescent Health (Add Health), a longitudinal study of more than 20,000 adolescents in 1995 who have been followed through adolescence and the transition to adulthood. Through her participation in the NICHD Family and Child Well-Being Research Network, Harris is studying the health status and health behavior of children in immigrant families and the role of social contexts in the acculturation of immigrant youth. Other current work is examining the determinants of nonmarital childbearing, including the impact of recent welfare reform policies. Harris was awarded the 2004 Clogg Award for Early Career Achievement from the Population Association of America. Among many professional commitments, she serves as elected vice president of the Population Association of America.

Aletha C. Huston is the Pricilla Pond Flawn Regents Professor of Child Development and the Associate Director of the Population Research Center at the University of Texas at Austin. She specializes in understanding the effects of poverty on children and the impact of child care and income support policies on children's development. She is a principal investigator in the New Hope Project, a study of the effects on children and families of parents' participation in a work-based program to reduce poverty, and collaborator in the Next Generation Project. She was a member of the MacArthur Network on Successful Pathways through Middle Childhood and an investigator for the National Institute of Child Health and Human Development Study of Early Child Care and Youth Development. Her books include *Children in Poverty: Child Development and Public Policy*; *Big World, Small Screen: The Role of Television in American Society*; and *Developmental Contexts in Middle Childhood: Bridges to Adolescence and Adulthood*. She is past president of the Division of Developmental Psychology of the American Psychological Association and President of the Society for Research in Child Development. She has received the Urie Bronfenbrenner Award for Lifetime Contributions to Developmental Psychology, the Nicholas Hobbs award for Research and Child Advocacy, and the SRCD award for contributions to Child Development and Public Policy.

Nancy Y. McGovern is the Assistant Director of a new research department within Instruction, Research, and Information Services (IRIS) at Cornell University. She is developing a service-oriented program to provide a focal point for research within Cornell University Library that aims to conduct research on a broad range of library topics through an ongoing program of funded and ad hoc research projects, apply research techniques to operational planning and projects, and produce a series of quality research products. She is also Digital Preservation Officer at Cornell University Library. She formulates preservation policy and serves as liaison to digital preservation projects and initiatives. She is also coeditor of *RLG DigiNews*, a bimonthly Web publication that focuses on digitization and preservation. Some of Nancy's professional presentations bear the following titles: "Virtual Remote Control for Web Resources," "'Commentary' on Archival Standards: A Snapshot of Our Professional Practices in 2004," and "Mapping Organizational Activities to the OAIS Reference Model." She is also involved in the following OAIS-related international development activities: a task force on Digital Repository Certification and a working group to develop OCLC/RLG Preservation Metadata Implementation Strategies.

Samuel L. Myers Jr. is Roy Wilkins Professor of Human Relations and Social Justice of the Hubert Humphrey School at the University of Minnesota and directs the Roy Wilkins Center for Human Relations and Social Justice. He specializes in the impacts of social policies on the poor. Myers has served as president of the Association of Public Policy Analysis and Management and was appointed to the Executive Council of the National Association of Schools of Public Administration. He has also served on the Association's policy council and on the American Economic Association's Committee on the Status of Minority Groups in the Economic Profession. Myers has consulted with the National Commission for Employment Policy, National Academy of Sciences, U.S. Civil Rights Commission, U.S. General Accounting Office, and U.S. Congressional Committee on the Judiciary, Subcommittee on Crime. He was on the academic advisory board of the National Forum for Black Public Administrators, National Council for Black Studies board of directors, and editorial boards of the *Journal of Policy Analysis and Management*, *Social Science Quarterly*, and the *Review of Black Political Economy*. In 1990, the *Review of Black Political Economy* recognized Myers as one of the top 20 U.S. Black economists.

Summer Training Program Offers New Courses

Registration is now open for the 2006 ICPSR Summer Program in Quantitative Methods of Social Research. The Summer Program Web site now contains a complete update for 2006, including a new time schedule, course descriptions, and other relevant information for those contemplating attending one or more of our Summer Program workshops. In addition to our core courses in research design, statistics, data analysis, and methodology, the Program offers the following new and noteworthy courses this year:

Four-Week Courses:

- Measurement of Race and Ethnicity
- Historical Demography

Three-to Five-Day Workshops:

- Aging & Health Among Latin Americans & Hispanics
- Statistical Analysis With Incomplete (Missing) Data
- Introduction to CrimeStat 3.0
- Spatial Data Analysis
- Project on Human Development in Chicago Neighborhoods
- People, Place, and Environment in Northeast Thailand
- Health Care Change in the U.S.
- Providing Data Services
- Child Care & Early Education Research Themes

Visit the Web site at www.icpsr.umich.edu/sumprog.

ICPSR
A PARTNER IN SOCIAL SCIENCE RESEARCH
INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

P.O. Box 1248
Ann Arbor, MI 48106-1248

Nonprofit Organization
U.S. Postage
PAID
Ann Arbor, Michigan
Permit No. 144

ADDRESS SERVICE REQUESTED

Moving? Please send us your new address along with your old mailing label.

50% recyled fiber content
20% post consumer recovered fiber