



Future Investments in Large-Scale Survey Data Access & Dissemination

**American National Election Survey (ANES)
General Social Survey (GSS), and
Panel Study of Income Dynamics PSID**

Principal Investigators Meeting
July 26 – July 27, 2010
Arlington, Virginia

Report prepared by:

Patricia White
Jan Stets
Regina Werum
Christina Farhart

Sociology Program
Directorate for Social, Behavioral and Economic Sciences
National Science Foundation

Acknowledgements

We wish to thank the many people who contributed to this project. Myron Gutmann, Assistant Director, Social, Behavioral and Economics Sciences (SBE); Judith Sunley, Deputy Assistant Director, SBE; Frank Scioli, Director, Division of Social and Economic Sciences, and the following program directors: Cheryl Eavey, (Methodology, Measurement and Statistics), Brian Humes (Political Science), Nancy Lutz (Economics), Carol Mershon (Political Science), Jacqueline (Meszaros, Innovation and Organizational Sciences), and Julia Lane (Science of Science and Innovation Policy) helped organize, conceptualize, and design the two-day meeting and participated in writing the report with additional help from Amy Friedlander (Senior Advisor). We also thank Allison Smith (Social and Political Sciences Cluster Program Specialist) for her assistance with meeting logistics. Finally, the success of the entire effort rests on the workshop participants who provided well-informed presentations, constructive comments, and insightful recommendations.

Workshop Participants

George Alter, Inter-university Consortium for Political and Social Research (ICPSR)

Peter Bearman, Columbia University

Andrew Beveridge, Queens College and the Graduate School and University Center at CUNY

Mark Chaves, Duke University

Matthew DeBell, Stanford University

Dan Benjamin, Cornell University

Darrell Donakowski, University of Michigan

Daniel Goroff, Office of Science and Technology Policy

Myron Gutmann, National Science Foundation

Pamela Herd, University of Wisconsin

David Howell, University of Michigan

Michael Hout, University of California at Berkeley

Brian Humes, National Science Foundation

Vincent Hutchings, University of Michigan

Simon Jackman, Stanford University

Nirmala Kannankutty, National Science Foundation

Dean Lillard, Cornell University

Samuel Lucas, University of California at Berkeley.

Nancy Lutz, National Science Foundation

Peter Marsden, Harvard University

Carol Mershon, National Science Foundation

Jack Meszaros, National Science Foundation

Timothy Mulcahy, National Opinion Research Center (NORC)

Steven Ruggles, University of Michigan

Narayan Sastry, University of Michigan

Frank Scioli, National Science Foundation

Allison Smith, Department of Homeland Security

Tom Smith, National Opinion Research Center (NORC)

Lynn Smith-Lovin, Duke University

Jan Stets, National Science Foundation

Jennifer Stoloff, Department of Housing and Urban Development

Lois Timms-Ferrara, Roper Center for Opinion Research

Regina Werum, National Science Foundation

Patricia White, National Science Foundation

Mark Wilhelm, Indiana University-Purdue University Indianapolis

Executive Summary

On July 26-27, 2010, The Division of Social and Economic Sciences (SES) of the National Science Foundation, Directorate of Social, Behavioral and Economic Sciences (SBE) convened a meeting of the principal investigators (PIs) of three major surveys at the Westin Hotel in Arlington, Virginia. The meeting, *Future Investments in Large-Scale Survey Data Access and Dissemination*, was organized by the Sociology Program with support from the Economic, Political Sciences, and Methodology, Measurement and Statistics Programs. The meeting was focused on access to American National Election Studies (ANES), General Social Survey (GSS) and the Panel Study of Income Dynamics (PSID) data, and represented part of a broad strategy across the social, behavioral and economic (SBE) sciences to link its programs to the priority of the federal government to make information, including data, more broadly accessible. The ANES, GSS, and PSID are long-standing pillars of the social science research enterprise and represent major investments in the data infrastructure by the SBE directorate. As SBE maintains and re-evaluates its investments in research infrastructure, it continuously seeks substantive and methodological innovation in how data resources are developed and disseminated. The meeting was centered on the three “gold standard” surveys, but also sought advice that could speak to broader infrastructure and data needs for the social and behavioral sciences. The goal of the meeting was to outline a vision for the future dissemination of data from the three surveys and other NSF-supported survey data collection efforts, taking into consideration data access, management, standards and stewardship, and resource requirements.

The twenty-four meeting presenters had wide ranging experience and expertise with data collection and dissemination. Along with the principal investigators (PIs) of the ANES, GSS, and PSID, presenters were directors of major data projects, which included the Integrated Public Use Microdata Series (IPUMS), National Longitudinal Study of Adolescent Health (Add Health), the Wisconsin Longitudinal Survey (WLS), the Comparative Study of Electoral Systems (CSES), and Cross-National Equivalent File study (CNEF); representatives of major data centers and repositories, i.e., Inter-university Consortium for Political and Social Research (ICPSR), Roper Center for Public Opinion, and the Data Enclave at the National Opinion Research Center (NORC); and experts from the user community.

The workshop was organized into four sessions to discuss current practices and promising opportunities for future data collection, access and dissemination. The four sessions were:

1. Data Access and Management for Large-Scale Data Resources: Stewardship, Standards, and Promising Directions.
2. The ANES, GSS, and PSID: Dissemination Tools, Services, Challenges, and Future Vision
3. The User Community: Accessibility, Usability, and Needs.
4. Broadening the User Community, Integration across Data Resources and Budgetary Requirements

First, representatives from major data survey projects, centers, archives and repositories briefly described their project or organization, their current access and dissemination practices, and the applicability of these activities to a larger dissemination platform. They were also asked to recommend optimal dissemination strategies for the three SBE-supported survey projects – ANES, GSS, and PSID – and similar data collections that might expand and enhance future data collection and dissemination. Second, the ANES, GSS and PSID PIs discussed their current dissemination tools and services along with challenges and how these might be addressed. Third, representatives of the user community discussed

data accessibility, usability, and future needs of the research community. They addressed questions of whether the ANES, GSS and PSID are serving the relevant and most comprehensive user community and the short- and long-term dissemination needs of this community. Finally, all attendees participated in a general discussion of the resources needed to improve the infrastructure for the management and dissemination of both raw data and research results emerging from these three large survey and other data capturing and collection efforts. At the end of the discussions, participants formulated five summary recommendations.

1. **Require that all future data and metadata be presented or documented according to a well-defined protocol.** This will allow desirable modes of data access, search, downloads, and documentation.
2. **Retrofit all legacy data and metadata (if it still exists) to become machine readable.** This will possibly open up vast amount of data for dissemination and analysis once issues of confidentiality and disclosure are resolved.
3. **Develop a federated archive portal to facilitate collaboration of best practices for data dissemination strategies for the ANES, GSS and PSID.** Such an archive might grant users access to ANES, GSS and PSID data from a single portal, rather than multiple portals with potentially different versions of the same data sets. The mechanism might also provide expert, intermediate (novice) and lay users access to all three main data sets from one site that, in turn, directs users to the project-specific website.
4. **Support the development of common online, open-source platform(s) or tool(s) to facilitate data search, downloads, and *basic* analysis, mostly for experts and novices (intermediate skill) users.** The tools should facilitate searching and downloading from one or multiple data sets at thematic levels and by variable; provide users with information about metadata, such as the provenance of the data; permit users to construct dynamic codebooks that reflect the actual data downloaded; and meet current data management standards.
5. **Develop formats and methods to disseminate actual data results to the general public and facilitate responsible “citizen science.”** Current dissemination efforts target expert users and the social science research community. A range of dissemination strategies includes ones to distribute actual results to the general public and encourage "citizen science" as a means of engaging the public in the conduct of social science research. These strategies would focus on deepening a public appreciation for the value and challenges of the science.

CONTENTS

Acknowledgements	1
Executive Summary	3
Background and Purpose	7
Synthesis of Discussion	9
ANES, GSS & PSID	
American National Election Studies (ANES).....	9
General Social Survey (GSS).....	9
Panel Study of Income Dynamics (PSID).....	10
Collection of Data	10
Dissemination of Data, Metadata, and Research Results	
Dissemination of Data.....	12
Dissemination of Metadata.....	14
Dissemination of Research Results.....	15
Recommendations	17
Appendices	
Appendix 1. Meeting Agenda.....	19
Appendix 2. Participants Presentations.....	23
Data Access and Management for Large-Scale Data Resources Stewardship, Standards, and Promising Directions	
George Alter, <i>Inter-University Consortium of Political and Social Research (ICPSR)</i>	25
Lois Timms-Ferrara, <i>The Roper Center for Public Opinion Research</i>	29
Steven Ruggles, <i>Integrated Public Use Microdata Series (IPUMS)</i>	33
Timothy Mulcahy, <i>Data Enclave, National Opinion Research Center</i>	37
Peter Bearman, <i>National Longitudinal Study of Adolescent Health</i>	39
Pamela Herd, <i>Wisconsin Longitudinal Study</i>	41
David Howell, <i>Comparative Study of Electoral Systems (CSES)</i>	43
Andrew Beveridge, <i>Social Explorer</i>	45
Nirmala Kannankutty, <i>The Science and Technology Enterprise, National Center for Science and Engineering, National Science Foundation</i>	47
PIs of ANES, GSS, and PSID: Dissemination Tools, Services, Challenges, and Future Vision	
Vincent Hutchings and Simon Jackman, <i>American National Election Studies (ANES)</i>	51
Peter Marsden and Tom Smith, <i>General Social Survey (GSS)</i>	55
Narayan Sastry, <i>Panel Study of Income Dynamics (PSID)</i>	59
The User Community: Accessibility, Usability, and Needs	
Mark Chaves, Duke University and Chair of the GSS Board of Overseers.....	63
Matthew DeBell & Darrell Donakowski.....	65
Michael Hout, University of California, Berkeley, and Principal Investigator, GSS.....	67
Dean Lillard, <i>Enhancing Data Dissemination: The Cross-National Equivalent File Study (CNEF)</i>	69
Samuel Lucas, <i>On Data Access, Usability, and User Community Needs</i>	73

Lynn Smith-Lovin, <i>Future Investments in Survey Data Access and Dissemination</i>	77
Jennifer Stoloff, <i>Potential for Non-Academic Use of PSID Data</i>	79
Mark Wilhelm, <i>Using ANES, GSS and PSID Data</i>	81
Appendix 3. Biographical Sketches of Presenters.....	83

Background and Purpose

In 2007, the National Science Foundation (NSF) held a workshop in which representatives of the social science research and data user communities discussed future directions for the General Social Survey (GSS). Among the recommendations outlined in the workshop report, *The GSS: The Next Decade and Beyond*,¹ was a call for greater data transparency, easier data access for the user community, and more user-friendly dissemination modes (especially an improved website and online search tools). In addition, the report called for the use of digital technologies to capture metadata and the development of cooperative relationships with the American National Election Studies (ANES) and the Panel Study of Income Dynamics (PSID) to share best practices for the incorporation of new technologies in data collection, access and dissemination. These recommendations also echoed suggestions expressed in the NSF *Cyberinfrastructure Vision for 21st Century Discovery*² report, published the same year.

Building on these reports, the Division of Social and Economic Sciences (SES) of the Directorate of Social, Behavioral and Economic Sciences (SBE) convened a meeting of the principal investigators (PIs) of the three major surveys the directorate supports. The meeting, *Future Investments in Large-Scale Survey Data Access and Dissemination*, took place over two days, July 26-27, 2010, at the Westin Hotel in Arlington, Virginia, and was organized by the Sociology Program with assistance from the Economics, Political Science, and Methodology, Measurement and Statistics Programs. The workshop is part of a broader strategy in the social and behavioral sciences to link to the priority of the federal government to make information, including data, more broadly accessible. The three surveys, American National Election Studies (ANES), General Social Survey (GSS) and the Panel Study of Income Dynamics (PSID), are long-standing pillars of the social science research enterprise and represent major investments in data infrastructure by the directorate. As SBE maintains and re-evaluates its investments in research infrastructure, it continuously seeks substantive and methodological innovation in the ways data resources are developed and disseminated. Although the meeting focused on the three so-called “gold standard” surveys, organizers also sought advice that could speak to broader infrastructure and data needs for the social and behavioral sciences. The goal of the meeting was to outline a vision for future data dissemination of the three surveys and other NSF-supported survey data collection efforts, taking into consideration data access, management, standards and stewardship, and resource requirements.

The meeting participants had wide-ranging experience and expertise with data collection and dissemination. (See Appendix 2 for biographical sketches.) They included the principal investigators (PIs) of the ANES, GSS, and PSID as well as directors of other major data projects (Integrated Public Use Microdata Series [IPUMS], National Longitudinal Study of Adolescent Health [Add Health], the Wisconsin Longitudinal Survey [WLS], the Comparative Studies of Electoral Systems [CSES], and Cross-National Equivalent File study [CNEF]); representatives of major data centers and repositories (Inter-university Consortium for Political and Social Research [ICPSR], Roper Center for Public Opinion, and the Data Enclave at National Opinion Research Center [NORC]), and experts from the user community. (See the Appendix 3 for a summary of the activities of the data projects, centers and repositories represented at the meeting).

¹ *The General Social Survey (GSS), The Next Decade and Beyond*, National Science Foundation Workshop on Planning for the Future of the GSS, NSF-0748, National Science Foundation, Arlington, VA.

² *Cyberinfrastructure Vision for 21st Century Discovery*, NSF 07-28, National Science Foundation, Arlington, VA.

The workshop was organized into the following four sessions.

1. Data Access and Management for Large-Scale Data Resources: Stewardship, Standards, and Promising Directions
2. The ANES, GSS, and PSID: Dissemination Tools, Services, Challenges, and Future Vision
3. The User Community: Accessibility, Usability, and Needs
4. Broadening the User Community, Integration across Data Resources and Budgetary Requirements

First, representatives of major survey data collections, centers, archives and repositories briefly described their project or organization, their current access and dissemination practices, and the applicability of these activities to a larger dissemination platform. They were also asked to recommend optimal dissemination strategies for the NSF-supported surveys and similar data collections that might expand and enhance future data collection and dissemination. Second, the ANES, GSS and PSID PIs discussed their current dissemination tools and services along with challenges and how these might be addressed. Third, representatives of the user community discussed data accessibility, usability, and future needs of the research community. They addressed questions of whether the ANES, GSS and PSID are serving the relevant and most comprehensive user community, and the short- and long-term dissemination needs of this community. Finally, all attendees participated in a general discussion of the resources needed to improve the infrastructure for the management and dissemination of both raw data and research results emerging from the three surveys and other data efforts. The group discussion yielded a series of recommendations outlined in this report. In addition to issues related to data dissemination, workshop discussion also pointed to several remaining obstacles to data dissemination as well as the need for future conversations regarding efforts to invest resources into migration from phone interviews to Internet and mobile devices as a means of data collection. Competing demands include maintaining a high response rate while considering cost containment strategies for indispensable personal interviews.

This report contains three major sections: Background and Purpose; Synthesis of the Group Discussion; and Recommendations. The meeting agenda, biographical sketches of meeting participants, and summaries of presenters' remarks are in the appendices.

Synthesis of Discussion

The ANES, GSS and PSID

The ANES, GSS and PSID are long-term survey projects that are building research infrastructure for the social and behavioral sciences. A major goal of the three survey projects is accessibility and ease of use that has "value-added" far beyond the original data collection effort. That value consists of providing access and tools that enable dissemination to wide ranging user communities -- from social scientists to advance knowledge and test theories, to teachers in secondary schools to explain basic statistical and analytic methods, to citizens outside of the higher education and research communities who use the data to generate basic descriptive statistics and graphs.

American National Election Studies (ANES)

The ANES is a 60-year time series of survey data collections that began in 1948. The ANES conducts national surveys of the American electorate in election years and carries out research and development work through pilot studies. The current ANES *has time-series* (i.e., the same questions are asked on successive surveys) election year data from 1948-2008. In presidential election years, the study is typically conducted both before and after the election (that is, a pre-election study and a post-election study), while for congressional election years the study has typically been conducted only after the election (a post-election study). *Pilot studies*, normally conducted in years when there is not a national election, are done to test new or refine existing instrumentation and study designs.

The ANES produces high quality data on voting, public opinion, and political participation over time to serve the research needs of social scientists, teachers, students, policy makers, and journalists who want to better understand the theoretical and empirical foundations of national election outcomes. To encourage maximum involvement from the user community, ANES developed The Online Commons. Through this vehicle, scholars around the world can participate in shaping the ANES by submitting ideas for survey questions and future topics and identifying issues in data collection. The ANES is a long-term participant in the Comparative Studies of Electoral Systems (CSES), an international collaborative program of research among election study teams that collect data on a common module of survey questions in their post-election studies, enabling analysis of U.S. elections in a comparative context.

The General Social Survey (GSS)

The GSS has provided data on contemporary American society since 1972, serving as a barometer of social change and trends in attitudes, behaviors, and attributes of the United States adult population. The GSS is a nationally representative personal interview survey of the United States adult population that collects data on a wide range of topics: behavioral items such as group membership and participation; personal psychological evaluations including measures of well-being, misanthropy, and life satisfaction; attitudinal questions on such public issues as crime and punishment, race relations, gender roles, and spending priorities; and demographic characteristics of respondents and their parents.

In 1984, the GSS stimulated cross-national research by collaborating with Australia, Britain, and Germany to develop data collection programs modeled on the GSS. This program of comparative cross-national research, called the *International Social Survey Program* (ISSP), now includes forty-eight

nations and enables researchers and analysts to place findings and trends from the United States within a comparative perspective.

The GSS website is extremely popular registering over 4,000,000 visits annually. The user community includes researchers, college teachers, university students, business planners, media and public officials. Sociologists, political scientists, economists, statisticians, survey methodologists, anthropologists, geographers, biologists, engineers, psychologists, criminologist, legal scholars, medical/health researchers, and business administration and management scholars all use GSS data, and its use is widely documented in publications.

The Panel Study of Income Dynamics (PSID)

The PSID is a longitudinal survey of a nationally representative sample of U.S. families that began in 1968. The PSID is the world's longest running nationally representative panel survey. With over forty years of data on the same families and their descendants, the PSID is considered a cornerstone of the data infrastructure for empirically-based social science research in the U.S. and the world. The long panel, genealogical links, and broad content of the data provide a unique opportunity to study evolution and change within the same families over a considerable time span.

The PSID collects information on the dynamics of human and social behavior. These data can be used to systematically investigate a myriad of questions in a variety of scientific disciplines involving the study of life-cycle opportunities and trajectories over time. The extended time series of data allows the estimation of robust, causal models and supports the study of economic behavior, for example, through changing conditions such as wage variations for different populations during the course of business cycles. In addition, the longitudinal data facilitate the conduct of cohort analysis as persons from one time period to another may be compared. These data also facilitate developmental analysis, as early experiences may be used to predict longer-term outcomes, such as the prediction of income and health in adulthood from early-life experiences. The long panel of data improves the precision of the measurement as multiple measures are collected within the same families as well as from multiple family members over a period of many decades.

Over 25,000 customized data sets have been downloaded from the PSID Internet-based data center. The website has tools to create customized data sets and codebooks including the intergenerational and family-based analytic files. Instructors in undergraduate institutions and secondary schools use data from the PSID to illustrate basic concepts in scientific methods, including the composition of data, its structure, and linkages among data, and to demonstrate statistical analysis to make inferences about human behavior.

Collection of Data

Common data formats are necessary for the wider accessibility of survey data. Future data collections by the ANES, GSS, and PSID must meet common standards for machine-readability. Most likely, data collection standards will be centered on the Data Documentation Initiative (DDI)-based schema³ and Extensible Markup Language (XML)⁴ format. However, DDI is one of several approaches, some of which are already followed internationally, that could be deployed. A decision

³ The Data Documentation Initiative (DDI) is an effort to create an international standard for describing data from the social, behavioral, and economic sciences.

⁴ Extensible Markup Language (XML) is a flexible text format that is used to exchange a wide variety of data from different sources.

would have to be made as to whether DDI is indeed the appropriate approach for describing social science data.

A decision about DDI is one of several technological choices that are looming. Future standards, while backwards compatible, must be designed to ensure a high level of “granularity” (detail) in the searchable metadata to enable investigators to establish links across sets of survey data or connections to external data sources in the biological and physical sciences, which are relevant to fields as diverse as epidemiology and urban studies. Consideration should be given to developing software tools to track publications and citations linked to the different data sets and to quantify users’ downloads of raw data as well as the processed or packaged data sets derived from the raw findings. Optimizing knowledge of the use of the three data sets will, in turn, help guide future data collection and dissemination efforts. However, substantial investments will be required to implement this time- and labor-intensive goal.

With respect to survey methodology, participants identified a need to complement traditional CAPI (computer assisted personal) and CATI (computer-assisted telephone) interview techniques with other approaches, for example, with Internet-based surveys. The ANES plans to include Internet-based surveys in addition to traditional face-to-face surveys in its 2012 wave, but no consistent procedure exists to integrate this new data collection approach into all three data sets. Thus, there is not a mechanism to assist in the long-term goal of migrating ANES, GSS, and PSID data collection from traditional to new technologies (including Internet and mobile devices). Future efforts to utilize these technologies need to maximize data that are collected initially in the same machine-readable format, regardless of whether they are obtained via CAPI, CATI, or other techniques. However, successful use of these technologies for data collection purposes will partly depend on factors outside of the PIs’ control – including the United States’ comparatively low high-speed Internet penetration rate relative to some other nations.⁵ Uneven access to broadband as a result of uneven infrastructure development may well affect survey projects’ ability to continue to collect uniformly high-quality data in a cost-effective inter-based approach above and beyond the three data sets featured in this report.

Dissemination of Data, Metadata, and Research Results

Data dissemination issues dominated the workshop discussion and focused on access and distribution of data, metadata and research results from the analysis of data and metadata. Participants framed the discussion of data dissemination using the following four questions.

1. Who constitutes the current user community, and how can it be expanded?
2. What are the best practices in data dissemination? How can these be improved?
3. What, if any, metadata dissemination practices exist? What is technologically feasible and desirable? What resources are needed to implement these new technologies?
4. What are the best practices in how results are disseminated to the wider public? How can dissemination of survey results be improved?

There is a broad range of users across the spectrum of surveys and data consortiums. The workshop participants identified three types of distinct users: experts (usually found in the academy), intermediate users (undergraduate students with some methodological training; some media and policy professionals), and lay users (most media and policy professionals, teachers, and the general public). Approaches to the dissemination of “raw” data (i.e., data that has not been processed for “official”

⁵ *Exploring the Digital Nation: Home Broadband Internet Adoption in the United States*, National Telecommunications and Information Administration, Washington, DC, November 2010.

release) o the first two groups continues to need improvement and may benefit from sending “data ambassadors” from the survey projects or expert data users to smaller, more specialist conferences. In addition, training workshops aimed at graduate students interested in using ANES, GSS, and PSID data, modeled after National Center for Education Statistics (NCES) training workshops would be a useful tactic. Great strides have been made in recent years to ensure maximum access to (public-use) data, and online tools have been developed to enable users to conduct searches and download data by variable lists, themes, and other modules (e.g., years, or sub-samples). However, no uniform online tool exists to facilitate data usage of all three big data sets. Rather, users must use different tools at the different sites, and federating or merging data from separate sources is extremely challenging.

Consensus emerged on levels of access for different user groups. At least for now, outreach to a broader user community should not include easier access to raw data for high school students or the general public. Data dissemination tools that are developed should improve lay person’s access to results (discussed below, under #4) and permit the construction of basic descriptive tables and graphs with public-use data. To train a larger next generation of data users, dissemination activities should include curriculum segments aimed at secondary school teachers, who may be interested in incorporating survey data (or results) into their social science classes. In particular, tools like the Roper Center iPOLL, which provides previously created visualizations of survey data patterns, might be useful. Similarly, online analysis tools currently offered by the Inter-university Consortium for Political and Social Research (ICPSR) and via Survey Documentation and Analysis (SDA) could be used to model future software development designed to facilitate basic online analysis of publically available data. Protocols should be created for granting the more sophisticated or “expert” users access to restricted data.

Dissemination of Data

Meeting participants approached the discussion of data dissemination by tackling issues of data access (public versus restricted), data platforms and mechanisms, online search tool, and standard data formats. The pros and cons of different options are offered, but issues remained unresolved. A discussion of each follows.

Public-Use versus Restricted Data Access

Data dissemination strategies will have to be multifaceted. Some data sets lend themselves to broad distribution to the media, policy analysts, and the lay public. These data sets tend to focus on the dissemination of aggregate-level data or completely anonymized public-use data. For example, GSS data are available via iPOLL at the Roper Center for Public Opinion Research⁶; also public access is granted, without embargo, to the entire repository of Comparative Study of Electoral Systems (CSES).⁷ Others data follow a more complex collection strategy and contain sensitive information that requires restricting access to some or much of the data to a smaller, quantitatively trained and largely academic community (e.g., PSID and Add Health). Efforts to add biomarkers to individual-level information and link participants and households to geospatial data will further increase the risk of disclosure associated with data commonly available on a restricted basis. To date, the ICPSR and Data Enclaves at NORC have played an important role in ensuring proper implementation of restricted data protocols, though new guidelines may need to be developed given the increasing complexity of data sets. This complex dissemination strategy will require innovation on several fronts.

⁶ The Roper Center iPOLL provides access to a database consisting of over 500,000 questions (including questions from the GSS) and topline marginals from US national samples, both current and historic (back to 1935), basic metadata, and links to data.

⁷ CSES is a collaborative program of research among election study teams from around the world.

Data Access through a Federated Archive Portal

Data dissemination strategies should be coordinated, so that users can access data from a single portal, rather than only through multiple portals with potentially different versions of the same data sets. The portal would include the ANES, GSS, and PSID websites as well as possibly other data access points (i.e., ICPSR, the Roper Center and SDA). The portal would not replace, but augment existing project-specific and other data access and dissemination sites. A *federated archive portal* would provide a mechanism whereby both expert and novice (intermediate group of users with a skill level between experts and lay users) could gain access to all three main data sets from one site that, in turn, directs users to the project-specific website (with persistent URL – universal resource locator – identifiers). Even though directed to project-specific sites, all three projects would use the same search tool, an open-source based interface that would have to be developed. This is an important change from current practice, where each data set has developed its own online search tool, but these are not compatible and thus impede data access to all but the most quantitatively inclined researchers and graduate students.

Future discussions beyond the current meeting would be needed to clarify exactly how the federated archive will operate, and how it can emulate best practices from similar ventures in the U.S. or abroad. While an online federated archive promises to broaden data access to a significantly larger number of expert and novice (intermediate skill) users, it will also affect the degree to which data are accessed via third-party providers. Moreover, this new approach will be ideally suited to disseminate aggregate and public-use data. Access to restricted data and the accompanying heightened risk of disclosure raises the issue of how to best balance data dissemination goals with confidentiality goals. In the existing model, researchers obtain access to restricted data through institutionally-specific IRB (institutional review board) protocols and contractual agreements with data providers. This approach puts archivists as gatekeepers at the center of the data dissemination process. If a federated archive is implemented where the data per se are placed at the center, and all constituents (archivists and users alike) are connected as if through spokes, how will requests for sensitive data dissemination be managed? This issue remains most pertinent to the PSID, GSS, and ANES (and by extension to similar data sets containing sensitive individual-level data, such as Add Health and the WLS).

Development of Open-Source Online Search Tool

Participants suggested that a top priority for the dissemination of data from the ANES, GSS, and PSID is the development of an open-source search tool for use with all three data sets. The ANES, GSS and PSID, however, should have online data dissemination strategies in place prior to working on the development of the tool. Such a tool will require substantial infrastructural investments, but soliciting expertise from multiple fields, including the computer science research community will assist in cost containment.

To initiate the development of an online open-source data search tool common to the ANES, GSS, and PSID, a workshop should be held to identify the relative advantages of commissioning tool development to either an academic or private-sector entity, or decide whether an existing commercial product (likely XML-based) is already available that would allow desirable modes of data access, search, downloads, and documentation. Workshop participants might include stakeholders from the business and academic communities.

The online search tool should contain several components. It should facilitate searching and downloading from one or multiple data sets at thematic levels and by variable lists. This requires development of syntax libraries applicable to variables and concepts across all three data sets. The tool should also provide users with information about metadata, such as the provenance of the data (e.g., actual question and sampling design); permit users to construct dynamic codebooks that reflect the actual

data downloaded, and meet data management standards likely to remain applicable well into the future. Ideally, the tool should enable survey project staff to track the frequency at which variables, modules, themes, and even metadata are downloaded. Such demand-side statistics may yield valuable information for the construction of future survey waves and questions.

Standardization of Data Formats

Increased standardization of formats is necessary to facilitate providing access to data for users with varying levels of data expertise. On the one hand, survey data currently are provided in several formats, typically associated with commonly used statistical packages used by social scientists (e.g., SAS, SPSS, and STATA). Providing data to users in several formats facilitates dissemination to a broader set of constituents (whose home institutions may not support a particular software package). On the other hand, there is merit in providing data in a single format. Multiple parallel data formats may impede data dissemination by delaying the release of the most current data and by increasing the risk of disseminating different versions of the data (just as providing data through multiple access sites does). It is unclear whether making data accessible in XML format will help circumvent these problems, and whether it is desirable that all future SBE-funded data collections conform to new, machine-readable data collection and dissemination standards.

Dissemination of Metadata

Metadata are commonly understood as “data about data”, and its dissemination may be the most complex problem discussed during the workshop. Metadata exist in many different formats (text and non-text-based). The records have also been stored in different ways depending on the date of the original data collection, the available technologies at that time (paper, scanned into pdf, and other formats), and access to storage facilities. Thus, metadata dissemination strategies must address two distinct sets of issues that bridge collection and dissemination: first, collecting and coding metadata associated with future waves of the ANES, GSS, and PSID surveys as collection and processing techniques evolve; and second, migrating (or “retrofitting”) metadata associated with earlier (i.e., legacy) surveys into formats and schema that are compatible with current and future collection efforts so that the full range of material is available for researchers in a coherent and consistent structure.

Dissemination of the metadata raises several issues that require further deliberation, and it remains unclear whether and how existing legal frameworks address issues of disclosure and human subject protection with respect to metadata collected by federally funded projects. Resolving these issues of user access to metadata, and the restrictions and conditions placed on such access is important because unrestricted access to metadata can mitigate processes put into place to ensure the protection of participants’ privacy. As such, open access to metadata could potentially increase the risk of disclosure and de-identification.

Calls for open access to data and metadata must be balanced against legal and ethical standards that, among other things, protect survey participants from inappropriate disclosure of personal information. It is not obvious whether all metadata will or should be released publicly following examples set by the Open Archives Initiative, which has its roots in efforts to enhance access to e-print archives as a means of enabling scholarly communication.⁸ Alternatively, only metadata associated with the formal content (the codebook or “paradata”) may be XML searchable and available to all users, but less structured information such as interviewer notes, handwritten and audio records may not. There is also the question of granularity or level of detail. While metadata associated with aggregated information may be made public or partially public, metadata associated with individual records is more sensitive and

⁸ Open Archives Initiative, <http://www.openarchives.org/OAI/OAI-organization.php>

access to personal identifiers would always be restricted. Thus, a new set of protocols will need to be developed that permit various levels of access to metadata, depending on the type of user (expert, intermediate, or lay public) and the scale of the underlying data that metadata record describes.

In addition to *who* can access metadata, the question about *where* users can access metadata needs to be addressed. The single portal envisioned for access to the federated archive might also provide links to respective metadata collections. Alternatively, each project-related website might direct users to data and metadata holdings. Most importantly, the search tool associated with metadata from the ANES, GSS, and PSID needs to be common to all applications.

Dissemination of Research Results

Improving the accessibility of results from federally sponsored projects is most pertinent for the general public, which includes the media, business community, congressional staff, teachers, and individual taxpayers. The term “citizen science” was used to describe the benefits of enabling the public to interpret and engage with data meaningfully. All data sets need to improve the public access to easily interpretable results. The ANES, GSS and PSID PIs generally acknowledged that data collection, quality and management continue to be priorities relative to the dissemination of data in raw and digested forms. Creating a federated archive with a common, open-source-based search tool might possibly be slowed down by lack of technical expertise, but the dissemination of results appears to be hampered by insufficient resources and available labor. It is not clear whether results would also be disseminated via the single portal associated with the federated archive, via the projects’ own websites, or via an NSF-based website.

The GSS website already makes it possible for lay users to construct basic uni- and bi-variate data presentations, but no uniform way exists for lay users to access prefabricated maps, charts, and tables, or construct basic data visualizations on their own. Although development of an online tool to access prefabricated graphs, maps, and tables could be an integral part of the tool designed to broaden access to actual data, it may be more efficient to draw on existing resources to aid in this aspect of dissemination. *The Social Explorer*⁹, *RoperExpress*¹⁰, Nesstar¹¹, and SDA¹² are good examples. For instance, *The Social Explorer* provides a model for providing lay users access to summary analyses and to a tool to create their own data visualizations, especially empirical patterns with geospatial features that are easily represented in maps. The *Social Explorer* was developed through a public-private collaboration involving the National Historical Geographic Information System (NHGIS) at the University of Minnesota Population Center, the Association of Religion Data Archives (ARDA), the New York Times, and GeoMicro, Inc. (which developed a well-known commercial geographic information system tool). Alternatively, iPOLL (pioneered by Roper) provides valuable insights into how cross-sectional and longitudinal opinion data trends can be easily presented in visual (rather than tabular) format. Nevertheless, PIs and other discussants expressed their concerns that existing tools could be challenging for new users and that the existing data sets required varying levels of statistical and social scientific literacy in order for these users to achieve legitimate results.

Attendees suggested the possibility of developing curriculum modules that incorporate the presentation of results from the ANES, GSS, and PSID. In this case, the challenge lies in synchronizing the social science modules with diverse state-level curriculum plans and standards. Future discussions are

⁹ See pages 45-46 of this report for a discussion of the *Social Explorer*.

¹⁰ See pages 29-22 for a discussion of *RoperExpress*.

¹¹ Nesstar is a software system for publishing data on the Web; see <http://www.nesstar.com/>.

¹² SDA (Survey Documentation and Analysis) is a set of programs used for the documentation and web-based analysis of survey data. SDA is developed and maintained at the University of California, Berkeley.

needed to articulate possible options for broadening dissemination of survey data as teaching and learning tools at the elementary and secondary school levels.

Recommendations

In the last session of the meeting, participants summarized their advice for investments in large-scale survey data access and dissemination. Five major recommendations were made, but participants also identified several obstacles to access and dissemination.

1. **Require that all future data and metadata be presented and documented according to a well-defined machine-readable protocol.** This will allow desirable modes of data access, search, downloads, and documentation.
2. **Retrofit all legacy data and metadata (if still existent) to become machine readable.** This will possibly open up vast amount of data for dissemination and analysis once issues of confidentiality and disclosure are resolved.
3. **Develop a federated archive portal to coordinate data dissemination strategies for the ANES, GSS and PSID.** Such an archive would grant users access to ANES, GSS, and PSID data from a single portal, rather than multiple portals with potentially different versions of the same data sets. The mechanism would also provide both expert and other users' access to all three main data sets from one site that, in turn, directs users to the project-specific website.
4. **Support the development of common online, open-source platform(s) or tool(s) to facilitate data search, downloads, and *basic* analysis, mostly for experts and intermediate users.** The tools should facilitate searching and downloading from one or multiple data sets at thematic levels and by variable; provide users with information about metadata, such as the provenance of the data (e.g., actual question and sampling design); permit users to construct dynamic codebooks that reflect the actual data downloaded, and meet current data management standards.
5. **Develop formats and methods to disseminate actual data results to the general public and facilitate responsible "citizen science."** Current dissemination efforts target expert users and the social science research community. A range of dissemination strategies includes ones to distribute actual results to the general public and encourage "citizen science." Data access and data dissemination strategies should move to target new or lay users to broaden the user community. Tools should range from one that allows the lay and intermediate users to produce basic descriptive statistics, tables and visuals to one that allow the expert to extract customized data files.

The meeting participants pointed out four major obstacles to data access and dissemination.

- **Data structure.** Some data are multilevel or have complex sampling designs that make the data extremely difficult for the non-expert to competently analyze without extensive training.
- **Data content.** Some surveys have sensitive data and meta-data that must be restricted to preclude the possibility of disclosure.
- **Data management.** Data access is granted via multiple providers with multiple versions and there are incompatibilities of variable names across waves.
- **Resources.** Facilitating data and metadata dissemination to a diverse set of users takes considerable resources. Data collection and coding tends to take precedence over long-term archiving and dissemination activities, especially when budgets are tenuous.

APPENDIX I.

Agenda

PI Meeting: ANES, GSS, and PSID

“Future Investments in Large-Scale Survey Data Access & Dissemination”

July 26 – July 27, 2010
Westin Hotel
Room: Hemingway 2 and 3
801 N. Glebe Road
Arlington, Virginia 22203

Monday, July 26

8:00 *Light Refreshments*

8:30 **Welcome and Overview of Meeting**

Myron Gutmann, Assistant Director, Social, Behavior and Economic Sciences (SBE)
Frank Scioli, Division Director, Social and Economic Sciences (SES)
Pat White, Acting Deputy Director, Social and Economic Sciences (SES)

9:00 **Session 1: Data Access and Management for Large-Scale Data Resources: Stewardship, Standards, and Promising Directions-** Moderator, Jan Stets, Program Director, Sociology

George Alter, Acting Director, Inter-university Consortium of Political and Social Research (ICPSR)

Lois Timms-Ferrara, Associate Director, Roper Center for Public Opinion Research

Timothy Mulcahy, Data Enclave Program Director, National Opinion Research Center (NORC)

Steven Ruggles, Principal Investigator, *Integrated Public Use Microdata Series* (IPUMS)

10:30 *Break*

10:45 **Discussion**

11:30 Peter Bearman, Co-Designer, *National Longitudinal Study of Adolescent Health* (Add Health)

Pamela Herd, Co-Principal Investigator, *Wisconsin Longitudinal Study*

12:15 *Lunch*

1:15 David Howell, Director of Studies, *Comparative Study of Electoral Systems* (CSES)

Andrew Beveridge, Professor of Sociology at Queens College and CUNY, Developer of *Social Explorer*

Nirmala Kannankutty, Senior Advisor/Senior Social Scientist, National Center for Science and Engineering Statistics (NCSES)

2:00 **Discussion**

2:45 *Break*

3:00 **Session 2: PIs of ANES, GSS, and PSID: Dissemination Tools, Services, Challenges, and Future Vision** – Moderator, Brian Humes, Program Director, Political Science

Vincent Hutchings and Simon Jackman, Principal Investigators, American National Election Studies (ANES)

3:30 **Discussion**

4:00 Peter Marsden and Tom Smith, Principal Investigators, General Social Survey (GSS)

4:30 **Discussion**

5:00 Narayan Sastry and Robert Schoeni, Principal Investigators, Panel Study of Income Dynamics (PSID)

5:30 **Discussion**

6:00 *Adjourn*

Tuesday, July 27

8:30 *Light Refreshments*

9:00 **Session 3: The User Community: Accessibility, Usability, and Needs-** Moderator, Nancy Lutz, Program Director, Economics

Mark Chaves, Professor of Sociology, Duke University, and Chair of the GSS Board of Overseers

Matthew DeBell, Director, ANES, Stanford University

Darrell Donakowski, Director of Studies, ANES

Michael Hout, Professor of Sociology, University of California, Berkeley, and Principal Investigator, GSS

Dean Lillard, Department of Policy Analysis and Management, Cornell University, and Senior Research Associate and Co-Director and Project Manager, Cross-National Equivalent File Study (CNEF)

10:00 *Break*

- 10:15** Samuel Lucas, Associate Professor of Sociology, University of California, Berkeley
- Lynn Smith-Lovin, Robert L. Wilson Professor of Arts and Sciences, Department of Sociology, Duke University
- Jennifer Stoloff, Social Science Analyst, Program Evaluation Division, Office of Policy Development and Research, Housing and Urban Development (HUD)
- Mark Wilhelm, Professor of Economics and Philanthropic Studies, Indiana University-Purdue University Indianapolis
- 11:00** **Discussion**
- 12:00** *Lunch*
- 1:00** **Session 4: General Discussion: Broadening the User Community, Integration across Data Resources, and Budgetary Requirements** – Moderator, Patricia White, Program Director, Sociology
- 2:00** Wrap-Up
- 2:30** *Adjourn*

APPENDIX 2. Meeting Presentations

Session 1: Data Access and Management for Large-Scale Data Resources: Stewardship, Standards, and Promising Directions-

George Alter, *Inter-university Consortium of Political and Social Research (ICPSR)*
Lois Timms-Ferrara, *The Roper Center for Public Opinion Research*
Timothy Mulcahy, *Data Enclave at the National Opinion Research Center*
Steven Ruggles, *Integrated Public Use Microdata Series (IPUMS)*
Peter Bearman, *National Longitudinal Study of Adolescent Health (Add Health)*
Pamela Herd, *Wisconsin Longitudinal Study*
David Howell, *Comparative Study of Electoral Systems (CSES)*
Andrew Beveridge, *Social Explorer*
Nirmala Kannankutty, *The Science and Technology Enterprise, National Center for Science and Engineering Statistics (NCSES)*

Session 2: PIs of ANES, GSS, and PSID: Dissemination Tools, Services, Challenges, and Future Vision –

Vincent Hutchings and Simon Jackman, *American National Election Studies (ANES)*
Peter Marsden and Tom Smith, *General Social Survey (GSS)*
Narayan Sastry, *Panel Study of Income Dynamics (PSID)*

Session 3: The User Community: Accessibility, Usability, and Needs

Mark Chaves, *Duke University and Chair of the GSS Board of Overseers*
Matthew DeBell, *ANES, Stanford University*
Darrell Donakowski, *ANES, University of Michigan*
Michael Hout, *University of California, Berkeley, and Principal Investigator, GSS*
Dean Lillard, *Cornell University, and Co-Director and Project Manager, Cross-National Equivalent File Study (CNEF)*
Samuel Lucas, *University of California at Berkeley*
Lynn Smith-Lovin, *Duke University*
Jennifer Stoloff, *Office of Policy Development and Research, Housing and Urban Development (HUD)*
Mark Wilhelm, *Indiana University-Purdue University Indianapolis and Center on Philanthropy Panel Study*

George Alter

University of Michigan
Acting Director

Inter-University Consortium of Political and Social Research (ICPSR)

The mission of the Inter-university Consortium of Political and Social Research (ICPSR) is to provide leadership and training in data access, curation, and methods of analysis for a diverse and expanding social science research community. The ICPSR was created in 1962 as a partnership among 21 universities to enable social scientists access to hard data. Initially, ICPSR was called Inter-university Consortium for "Political" Research since the primary focus was political science data. The ICPSR has a wider scope, has grown tremendously, and currently has almost 700 members nationwide. The ICPSR archive holds data that span the social and behavioral sciences with data available instantaneously, 24 hours a day, seven days a week. It has summer program in Quantitative Methods that attracts 800 participants, and an *Online Learning Center* and *TeachingWithData.org* that promotes quantitative literacy in undergraduate teaching.

Data Access and Users

ICPSR membership is comprised of universities, government agencies, and other organizations such as libraries and research institutes. One key aspect of membership is that there is an "official representative" at every member campus. Faculty, staff, and students of member institutions have full direct access to the data archive and to all of ICPSR analysis tools and technical support. ICPSR offers public access to certain data. These include data in the *Topical Archives* that pertain to criminal justice, health and aging, substance abuse and mental health, child care, and health and medical care (see Chart 1) Access to restricted data (i.e., where confidentiality is an issue) is granted through a data use agreement and the onsite Data Enclave.

Chart 1. Inter-university Consortium for Political and Social Research (ICPSR) Topical Archives

Topics of Archives

- Child Care and Early Education Research Connections
- Data Sharing for Demographic Research
- Health and Medical Care Archive
- National Addiction & HIV Data Archive Program
- National Archive of Computerized Data on Aging
- National Archive of Criminal Justice Data
- Resource Center for Minority Data
- Substance Abuse & mental Health Data Archive

Currently, there are 7,500 data collections in the archive, and it grows by 300-400 collections each year. Data are from several sources that include depositors, funding agency mandates, replication data sets, expert recommendations, series collections, new data combinations, and digitization and data entry. Thus, ICPSR is not just data. The *ICPSR Bibliography* links publications to data, has more than

56,000 citations to books, journal articles, dissertations, and other papers and publications and has the full text available for many publications.

College students are the greatest users of ICPSR data; approximately 35 percent of ICPSR users are graduate students and approximately 29 percent are undergraduates. University faculty represents 17 percent of ICPSR users. The *ICPSR Bibliography* is a popular service for data related literature in that it links publications to data. ICPSR recently launched an Online Learning Center initiative that takes a data set and puts it into a lesson plan. The intent is to get professors to integrate data into their teaching by making the integration easy. This practice has been expanded to a new website, TeachingWithData.org, which provides data modules from the Quantitative Social Science Data Library. The Online Training also includes a *Digital Preservation* website and tutorial. Chart 2 shows the different user communities, their level of access, and types of tools they require.

Chart 2. Inter-university Consortium Political and Social Research (ICPSR) User Sectors

Sector	Types of Tools	Types of Access
Academic - Research	Raw data	Full access
<u>Academic - Teaching</u>	<u>Teaching modules and lesson plans; slices of the data</u>	<u>Access through one source like SDA or ICPSR</u>
<u>Business, Media, Nonprofits</u>	<u>Depending upon experience with data</u>	<u>Easy access and explanations with meaningful results</u>
Government	Depending upon experience with data and what the data is being used for.	Easy access and explanations with meaningful/policy relevant results

Preparing for the Future User

Data users want the capability to do a single search for multiple variables and complete simple data extractions. They also want help with complex data management and analysis, extracting data, and merging geo-coded contextual data from other sources. Users also want the ability to link to publications using the same data. Provision of this type of capacity caters to the graduate student and less sophisticated user. These classes of users generally want a configuration that alleviates the need to merge data from other sources, which can be very complicated. They want access, like the *Web of Science* (an online academic citation index that provides access to multiple databases), to other sources using the same databases. Many students, scholars and the public get access to data from the General Social Survey, American National Election Survey and Panel Study of Income Dynamics through ICPSR, despite it not being the original point of contact.

“Idea” Data Access and Use

Users could be better served by having easy access to a broad range of data and metadata. Data standards could be applied across multiple data resources to facilitate more efficient data harvesting, compilation and analysis. Chart 3 outlines major “idea” requirements for enhanced data access and use..

Chart 3. “Ideas” for Data Access and Use

<p>Federated archives. Create federated data archives. Currently, ICPSR has a shared data catalogue with the Data Preservation Alliance for the Social Sciences (Data-PASS) and the Council of European Social Science Data Archives (CESSDA) is working on a set of projects to federate the catalogs.</p>
<p>Metadata. Metadata would be in machine actionable (xml) format and have variable level descriptions, a taxonomy/thesaurus, and provenance (i.e., question text, universe and source variables to provide information about from where the variables came).</p>
<p>Standards. Existing standards are available via the Data Documentation Initiative (DDI). The Open Archives Initiative - Protocol for Metadata Harvesting could be used to share data across archives. The direct transfer of metadata from producers could be done using the Michigan Questionnaire Documentation System, which transfers directly into DDI and captures many aspects of provenance. Newly emerging companies such as Collectica have the capacity to compile and analyze metadata.</p>
<p>User Tools. Access and analytic power would be increased through the development and use of dynamic codebooks (linked to the data extraction and to show how they're linked to one another). Data extraction capability, harmonization abilities, and tools to create syntax libraries (libraries of the command files that capture what users have done doing the data merging tasks) would be part of the user toolkit. Merging data sets would be made easier by having both in DDI and there would be an ability to create merged codebooks.</p>
<p>Data Citation Harvesting. Making available a consistent citations format in journalism and having persistent identifiers, digital object identifiers (doi) or a permanent URL, and handles would enable citation harvestings. The data uses would be cited, which gives recognition to PIs who collect the data. Currently, data citations within journals are inconsistent and not enforced and collecting the bibliography is very expensive. The use of doi would make it easier for organizations to identify and make a specific data set accessible. This would provide the needed permanent identifiers.</p>

Lois Timms-Ferrara
 University of Connecticut
 Associate Director
Roper Center for Public Opinion Research

The Roper Center was created in 1940s after World War II by Elmo Roper and George Gallup. The Roper vision was to archive and preserve Gallup polling data for “historians,” providing them with the context in which to study the events of the period. That is, poll data would provide an invaluable measure of public attitudes or thoughts during particular social, economic, political, national and international events. Polling data have proliferation since the 1940s and are valuable social science research tools. The Roper archives are a fairly comprehensive collection of public opinion surveys on policy matters.

The Center is a 501(c)(3) non-profit (or "charitable") organization governed by a board of 25 directors who are a combination of survey practitioners who produce the data in the archives, academic and commercial users, and information technology specialists. The Center is financed primarily by user fees, a few small grants and until now its current home, The University of Connecticut (UConn). UConn has generously supported the archiving functions of the Center over the last 30+ years, but in the current economic climate the Center is looking at different vehicles for sustainability. This includes options that involve different ways of disseminating data

The Center archives two forms of survey data—*Individual Data set Files* (ASCII and SPSS data formats) and *Reports and Releases*. It also sometimes archives supplemental materials for data sets such as the analysis provided by the survey organization or sponsor as well as reports from survey organizations that archive their data sets elsewhere. Most Roper Center holdings are from the commercial and media sectors. These holdings tend to be smaller commercial polls, typical with 60-90 variables and 1000-3000 respondents. The Center archives between 400 and 600 new studies each year. Poll data are deposited within a few months to a few years after collection. Data depositors do not provide financial support with their data. Also, over 7,000 surveys conducted in other countries are archived at the Center, including nearly 3,000 from Britain, and more than 1,300 Latin American surveys. Multi-country collections are among the more current of the international holdings.

Chart 1. Roper Center Holdings: US National and State Polls

The Gallup Organization	National and State Exit Polls
Pew Research Center	<i>The Los Angeles Times</i>
<i>Newsweek</i> magazine	Program on International Public Attitudes
Associated Press	<i>Time</i> magazine
General Social Survey	National Opinion Research Center
CNN	CBS News/The New York Times
Public Agenda Foundation	Abt/SRBI
ABC News/ <i>The Washington Post</i>	<i>USA Today</i>
Kaiser Family Foundation	Opinion Research Corporation
	and many others...

The User Community

The Roper Center user community includes the academic, non-profit, government, and private sectors and the news media. Its membership has grown from 39 to 200 during the period of 2003-2009. Most users are academic institutions, with users from the social science, business, public health, and journalism communities leading this growth. Only about half of the schools that use Roper data are research-intensive; there are a growing number of bachelors-level schools, and recently the user community is broadening to community colleges.

The latest market that the Center is engaging is high schools. A Center summer initiative involves developing classroom tools and training high school teachers to integrate data resources in civics, social studies, history, economics, language arts and mathematics classes. The most important lesson learned is that tutorials have to be very detailed and connect to specific curricular areas. This effort has moved the Center to consider licensing with third party distributors and is embracing the opportunities that social networking presents. The Roper Center also has members from non-profit special interest groups that include think tanks and political parties, but has received only a small amount of interest from the government sector.

The private sector comprised of media and corporate affairs offices, market researchers, public relations, and the financial industry demands different kinds of access tools than the research and academic communities. The private sector is also demanding and their need is immediate. The DATA MUST BE FRESH. It should be noted that the media does not pay for the conducting research, and corporate memberships come and go, and are tied to specific projects and project managers.

The Center disseminates 30,000 data sets annually. Seventy-five percent of the US collection is available for direct download. Missing collections are primarily older surveys conducted outside of the US. On-demand processing is available to members. The two major access tools are *RoperExpress* for data files and iPOLL for questions. *RoperExpress* is a catalog search tool, with the ability to provide data documentation, data set download, and tabulations. The Center disseminates 9,500 US national polls, 1,500 state polls and 7,000 non-US polls via *RoperExpress*. It provides an immediate download and tabulations. The iPOLL database consists of over 500,000 questions and topline marginals from US national samples, both current and historic (back to 1935), basic metadata, and links to data. About 400-500 non-demographic variables are added to iPOLL weekly and over 43,000 questions are downloaded on average per month.

Chart 2. The Roper Center Access Tools

Access Tools

Data: RoperExpress

- Catalog search tool
- Documentation
- Dataset Download
- Tabulations

Questions: iPOLL

- 500K Toplines
- US National Samples
- Current and Historic
- Basic Metadata
- Links to Data

[Exploratory Data Analysis](#)
SDA Study-level Modules
Subgroup banners within iPOLL

FOR PUBLIC OPINION RESEARCH
Roper
CENTER

CHOICE

July 26, 2010 6

Roper data from reports, press releases and survey data sets are recognized as being "the fastest source" as well as easiest and safest way to access this information. It was named by *Choice* as an Outstanding Academic Title in 2009. Roper begun an initiative, Exploratory Data Analysis, in spring 2010 that is data release and data set based. It provides tools for exploratory data analysis, prepares SDA (Survey Documentation and Analysis) study-level modules and subgroup banners within iPOLL to offer basic subgroups of data. This effort is to be very responsive to all of the user communities, making it possible to complete secondary analysis ranging from the calculation of simple cross-tabulations to more complex SDA analysis. By fall 2010 the Center anticipates it will make 2,500 studies available in this format.

Broadening Access and Optimal Dissemination

The value of iPOLL-like capabilities is that they provide granular search capability on the variable level, a synthesis of results, data set files that are linked, and links to detailed metadata. Currently, 10,000 questions from the GSS data are included in iPOLL. Plans are to integrate core variables from the American National Election Survey (ANES). The iPOLL *Topics at a Glance* provide results from predetermined searches on broad subjects. A few recent questions and results are displayed with each search. Any search provides access to historical data on the topic, links to related data sets, and this information is available to all users. This type of view into the iPOLL database allows users to see others' data. For example, it is possible to view five different Pew Research Center websites, the Kaiser Family Foundation or CBS News websites.

Strategic thinking around dissemination and access must address questions of how to effectively reach the various audiences and new audiences, and how to respond to the new stake holders. Specifically, who do you want to reach? Are there new audiences? What are the needs specific to those groups? How do we assess correlations across user communities? How do we evaluate needs of all stakeholders?

Needs for Optimal Data Access

There are several general factors that would enhance data access and dissemination. These include having a common metadata documentation standards, flexible platforms working, and with archives and collaborators. The Roper Centers suggests the following are needed for optimal data access.

1. **Common Metadata and DDI (Data Documentation Initiative) Compliance.** Require compliance with established DDI standard, self-documenting extracts, common transport format, provenance or link back to the source, and an enabling of more universally written applications, visualizations and analytical tools.
2. **Involvement of archives earlier in the data life cycle.** Release of data and documentation sooner and provide improved support for dissemination and use.
3. **Consideration of flexible platforms.** Partners with users and to enable public-driven science and citizen science.
4. **Collaboration with various partners.** Investigate third party distribution for particular elements of data access and documentation and integrate services.
5. **Organize and standardize a structure for usability.** This include developing a package for usability with attention to user groups needs, providing multiple access points, and permitting access, and enabling the most granular levels possible.

Steven Ruggles
University of Minnesota
Principal Investigator
Integrated Public Use Microdata Series (IPUMS)

The Integrated Public Use Microdata Series (IPUMS) project is building a platform for the dissemination of Census data. Prior to IPUMS it was very hard to use United States Census micro-data. IPUMS created micro-data samples (see Chart 1) that granted users access to Census data outside the Census Bureau. Analyses completed using these data created a big flurry of publications. The IPUMS project was to produce integrated data sets for selected years with one codebook, harmonized codes, and integrated documentation.

Chart 1. Public Use Microdata Samples of the Census in the United States, 1964-1990

<p>1964 - 0.1% sample of the 1960 Census</p> <p>1972 - six 1% samples of the 1970 Census</p> <p>1973 - 1% sample of the 1960 Census, harmonized with 1970, led to a burst of cross-temporal research</p> <p>1982 - 5% and 1% samples of 1980 censuses. Incompatible coding, format, and documentation compared with earlier censuses</p> <p>1980-1990 - historical samples of censuses from 1880, 1900, 1910, 1940, and 1950 produced at three universities and the Census Bureau, each with unique coding, format, and documentation</p>
--

In 1991 through an NSF-supported IPUMS award, integrated data sets for 1850, 1880, 1900, 1910, and 1940-1990 were produced with harmonized codes and integrated documentation. The preliminary version was released in 1993 and the first website was available in 1994. The goal was to make it so that users could make their own census data extract request instead of asking the Census Bureau to do it for them. By December 1995, IPUMS had the first interactive web-based data extract system. In 1996 IPUMS had a fully automated online registration and the ability to extract web pages generated dynamically based on prior user selections. By 1997 hypertext variable-level documentation from every stage of the extract system was accessible. With this, users can now create their own variables and the IPUMS project keeps the command files forever.

Between 1992 and 2010, IPUMS USA have added all other census years, the American Community Survey (ACS) and Current Population Survey (CPS) and expanded decennial samples. The database has grown to 160 million microdata records spanning 1850 to 2009.

IPUMS-International

Creating IPUMS-International to harmonize census data from different countries created new challenges. In 1999, the US census IPUMS model was expanded to the rest of the world to preserve, integrate and disseminate the data. Now IPUMS has data for 93 countries and 160 censuses are currently being disseminated. This volume of data creates a difficulty for dissemination. There is just too much information so there has to be multiple filters. The documentation can be brought up for any variable, but it is filtered by the data search. The enumeration text for all censuses selected can be viewed as well as the questions on the original census forms. Data dissemination is built on a shopping cart model where countries and variables of interests are selected and put in a cart for “check out.” IPUMS disseminate about a terabyte of data each week. The Minnesota Population Center is currently working on the American Time Use Survey Data Extract Builder and National Historical Geographic Information System. It will unveil a new dissemination system in January 2011 to facilitate the increasing user traffic.

Key Dissemination Needs

To enhance and improve data dissemination, generalized open source software tools for metadata creation and management, semantic integration, and tools for the dissemination of these (this is where most funding goes) are needed. Once these tools are developed they can be shared from project to project. The IPUMS project has tried to create a more accessible longitudinal file, but the metadata is inadequate. The tools to create the longitudinal metadata file would address the need to have rich DDI (Data Documentation Initiative) for data sets across time and across data sets to make the compatibility possible. Finally, flexible tools that can be customized for each data set are needed to both enhance data access and economic efficiency, given limited resources.

Diagram 1. IPUMS- International Participation

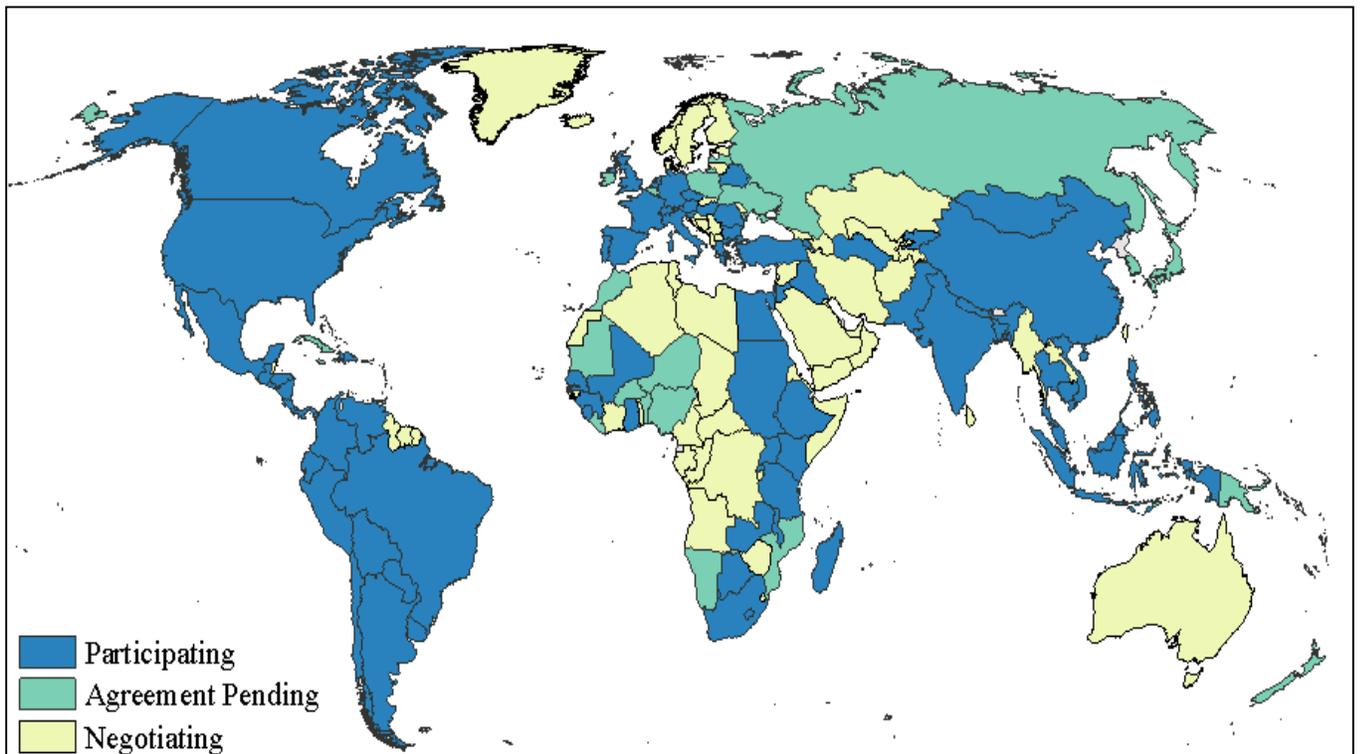


Diagram 2. Description of IPUMS Data Samples

MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA



[Home](#) | [Variables](#) | [Create Extract](#) | [FAQ](#) | [Contact Us](#) | [Login](#)

IPUMS Sample Information

Argentina	1970-1980-1991-2001	Hungary	1970-1980-1990-2001	Philippines	1990-1995-2000
Armenia	2001	India	1983-1987-1993-1999-2004	Portugal	1981-1991-2001
Austria	1971-1981-1991-2001	Iraq	1997	Puerto Rico	1970-1980-1990-2000-2005
Belarus	1999	Israel	1972-1983-1995	Romania	1977-1992-2002
Bolivia	1976-1992-2001	Italy	2001	Rwanda	1991-2002
Brazil	1960-1970-1980-1991-2000	Jordan	2004	Saint Lucia	1980-1991
Cambodia	1998	Kenya	1989-1999	Senegal	1988-2002
Canada	1971-1981-1991-2001	Kyrgyz Republic	1999	Slovenia	2002
Chile	1960-1970-1982-1992-2002	Malaysia	1970-1980-1991-2000	South Africa	1996-2001-2007
China	1982-1990	Mali	1987-1998	Spain	1981-1991-2001
Colombia	1964-1973-1985-1993-2005	Mexico	1960-1970-1990-1995-2000-2005	Switzerland	1970-1980-1990-2000
Costa Rica	1963-1973-1984-2000	Mongolia	1989-2000	Tanzania	1988-2002
Cuba	2002	Nepal	2001	Thailand	1970-1980-1990-2000
Ecuador	1962-1974-1982-1990-2001	Netherlands	1960-1971-2001	Uganda	1991-2002
Egypt	1996	Pakistan	1973-1981-1998	United Kingdom	1991-2001
France	1962-1968-1975-1982-1990-1999	Palestine	1997	United States	1960-1970-1980-1990-2000-2005
Ghana	2000	Panama	1960-1970-1980-1990-2000	Venezuela	1971-1981-1990-2001
Greece	1971-1981-1991-2001	Peru	1993-2007	Vietnam	1989-1999
Guinea	1983-1996				

Sample design	Sample fraction (%)	Households	Persons	Weighted	De jure De facto	Census date (d-m-yr)	Smallest geography	Collective dwellings	Notes
Argentina 1970	2	129,728	466,892	no	de facto	30-09-70	department	yes	
Argentina 1980	10	672,062	2,667,714	yes	de facto	22-10-80	department	yes	
Argentina 1991	10	1,148,351	4,143,727	yes*	de facto	19-05-91	department	yes	Missing data for several key variables requires use of alternative weight variable*
Argentina 2001	10	1,040,852	3,626,103	no	de facto	17/18-11-01*	department	yes	
Armenia 2001	10	81,929	326,560	no	de jure	09-11-01	province	no	
Austria 1971	10	264,655	749,894	no	de jure	12-05-71	NUTS3 region	yes	
Austria 1981	10	283,693	756,556	no	de jure	12-05-81	NUTS3 region	yes	
Austria 1991	10	310,099	780,512	no	de jure	15-05-91	NUTS3 region	yes	
Austria 2001	10	341,035	803,471	no	de jure	15-05-01	NUTS3 region	yes	
Belarus 1999	10	385,508	990,706	no	de facto	16-02-99	region	no	
Bolivia 1976	10	121,378	461,699	no	de facto	10-07-76	province	yes	
Bolivia 1992	10	177,926	642,368	no	de facto	03-06-92	province	yes	
Bolivia 2001	10	239,475	827,692	no	de facto	01-09-01	province	yes	

Timothy Mulcahy
National Opinion Research Center (NORC)
Director
NORC Data Enclave

Current data dissemination methods are through the use of public use files, remote access, data extracts, and tabulation engines that are generally slow. There are issues with this approach especially the quality and timeliness of access to public use files, security concerns with licensing and distribution, and the cost and accessibility of physical research data centers. The National Opinion Research Center (NORC) Data Enclave major focus is on how to provide secure access to microdata. Thus, the Enclave has created a secure environment for accessing sensitive data. It provides access through remote desktop, encryption, and audit logs. Security is enabled by a controlled information flow and group isolation. Finally, the Enclave is a facility for software and tools. These select tools include Stata, SAS, SPSS, matlab, R, Mplus, nlogit, Microsoft 2007, LISREL, adobe PDF reader, and the IHSN micro data management toolkit.

The mission of the Data Enclave is to promote access to sensitive micro data, protect confidentiality (portfolio approach), and archive, index and curate micro data. It also encourages researcher collaboration via a virtual environment of discussion forums, wikis, blogs, and instant messaging. The Enclave sponsors are the Departments of Commerce and Agriculture, National Science Foundation, Annie E. Casey Foundation, and the Center for Medicare and Medicaid Services. Today the Data Enclave has approximately 200 active researchers and 500 on the enclave listserv. The Enclave is increasingly accepted as a model of providing secure remote access to sensitive data. An emphasis on building and sustaining virtual organizations or laboratories has provided it the opportunity to build upon the power of metadata. Its major contributions are in helping with disclosure analysis and automating processes and in developing customized metadata driven tools.

The Enclave is changing the way we look at classic dissemination in a digital age with a survey life cycle by using a more fitting dynamic feedback model. It leverages the power of the metadata that we use every day that all use the same XML. The Enclave producer portal is an information source, providing general info, background info, announcements, and a calendar of events about the topic of the week. The enables knowledge sharing through discussion groups, wiki shared libraries of metadata/reports, scripts, and research papers. The portal has user support that includes *Frequently Asked Questions* and technical support and makes content fully editable by producers and researchers using a simple web-based interface. The Enclave facilitates private research group portals with similar functionalities, and it has as a customized researcher platform.

The Enclave employs innovation in its dissemination process. It has a disclosure control strategy that optimizes management processes for archiving and disclosure review. It facilitates workflow for the review and export of output requests to get a robust data interface across a whole research group. With researchers who agree to take their code and run it through Lucene index, the Enclave can give feedback on what variables are being used and which need to be expanded.

The only vulnerability with a secure remote access such as the Data Enclave is that the researcher is accessing data from their own machine. The Enclave cannot update the user's machine, so there is no way to know where it has been. What can be done to reduce the risk quotient? It could be reduced by sending out machines as access nodes directly into the NORC Data Enclave. Data producers have put together different types of secure machines, and not every data producer is the same. So, there needs to be flexibility in how secure access is provided. The Enclave is conducting a pilot test with five universities to monitor real data access and push updates. The machine can be a safe access node for any data set.

Peter Bearman
Columbia University
Professor of Sociology and Co-Designer
National Longitudinal Study of Adolescent Health (Add Health)

The goal of the National Longitudinal Study of Adolescent Health (Add Health) design was to collect data on independent variables with the hope that data users would discover dependent variables of which the designers had never thought. This led to a concentrate first on the getting the context in which adolescents were embedded right, and then on exciting saturation samples, twin design, genetic samples and so on in order that the study would grow. Add Health shares a similarity with the Panel Study of Income Dynamics (PSID) in that it seeks to make sense of the lives of adolescents and had to embed this deeply into the social context. We struggle with the sensitivity of the data collected.

The design of Add Health started with schools and interviewed all students at the school. This created many issues with respect to dissemination and deductive disclosure. These features of the data has made is hard to use. The clustered designs make deductive disclosure trivial if you know someone in the survey; it is not hard to find someone in the data set. The Add Health team is concerned about people, such as irate parent who would be motivated to find a sexual partner of their child. Thousands of people use parts of the Add Health data structure. For example, over 5,400 researchers have used the data. Add Health expanded its user base by organizing user conferences. With support from the NIH (National Institutes of Health) Add Health held user conferences. With such a complicated data set, most of the users tended to be younger scholars and conferences created an environment for them to give papers and learn from one another. This has built a user community with the bi-products of networking and a forum for research. User conferences are held at the annual meetings of the American Sociological Association, American Public Health conferences, and at other locations. At these, marketing of the survey data is done to attract young users. There is a very strong advantage in learning the data structure and the users conferences help facilitate that training.

Add Health also provides a "baby" public use data set designed to get people (especially students) quick access to data. The data set is drawn from the core sample and information such as the id numbers of friends or siblings removed, so that the data cannot be linked. The baby public use data set did not work the way we thought it would, for training on data use. We found that people tried to use the "baby" sample data for publishing articles.

The clustered sample design led to the need to help researchers correct for design effects in their estimation. Specifically, we have to set standards for getting around design effects. Meanwhile, standard analysis strategies were propagated to ensure that studies could speak to one another. For example, we simplified how to compute race and age for the users. Unusual elements of the data structure are described in publications made available by Add Health, and a number of user guides have been developed and are made available.

It should be noted that restricted data required an IRB (Institutional Review Board) approval. Initially this requirement was quite novel, but is now quite routine. Individual researchers had to be associated with a unit with an IRB and the unit had to validate that the systems were safe to secure IRB approval before they were granted access to the data. Every user file had a hidden signature that identified a purchaser. This signature was to track data use and make sure the data sets did not just get passed around. The IRB approval process has had both an up and a down side. The upside is that groups of universities have been disciplined for violating the data access terms. The downside is that IRBs have the idea that a single machine physically locked in a room with a password is safer than a data enclave. Data contracts are now handled out of ICPSR (Inter-university Consortium for Political and Social Research). Prior to this relationship with ICPSR, one person did all of the data dissemination.

Lessons Learned

We tried the idea of a public use baby data set that would be useful as a training tool. We, however, found that this was not very useful in that serious articles could not be developed based on analyses from the sample. Add Health data are still difficult to access, particularly the sexual relationship data. We also found that disciplining IRBs has been difficult. Finally, disclosure of Add Health data is an ongoing issue. With each addition to the Add Health data, the life course data sets become more explosive over time. This is because as some of the sample advances through the life course to positions of power and high status jobs, their behavior becomes more a subject of interest.

Pamela Herd
University of Wisconsin
Co-Principal Investigator
Wisconsin Longitudinal Study (WLS)

There will be some changes over the next few years in the Wisconsin Longitudinal Study (WLS). This presentation provides an introduction to the WLS, and discusses data content as a dissemination strategy, data accessibility, general promotion and future plans. The latter includes two broad plans of expanding the users of the data and collecting data on health and aging in late life.

The WLS started with the "Happy Days" cohort, Wisconsin High School graduates from 1957. The original sample size was 10,317. Looking for a state that is a microcosm of the whole country? You will not find it in Iowa or New Hampshire, sites of first presidential primaries. There are 25 states that come closer to average statewide measures on important characteristics such as race and income. The Badger State comes closer than any other to state-by-state averages on 12 key measures, according to a 2006 analysis by CNN Polling Director Keating Holland that takes a fresh look at U.S. Census data.

There have been seven surveys completed of the original respondents, parents and siblings, with the most recent in 2010 (see Chart 1 below). The 1957 survey collected data on social background, high school courses, social influences (i.e., teachers, parents, and peers), educational and occupational plans, military and marriage plans and parental support for college expenses. Each survey collects measures of important aspects of the sample's life course. The WLS collects both socio-demographic and administrative data. The socio-demographic data domains include social/family, education, employment, job characteristics, marital history, children (non-normative), physical/mental health, income and wealth, retirement and pensions, cognitive performance, and leisure time activities. Administrative data is comprised of test scores, class rank, parents, DNA, yearbooks, national death index, and soon, social security earnings and benefit histories. The relational structure of the data in the WLS makes it possible to do some network data (see Diagram 1).

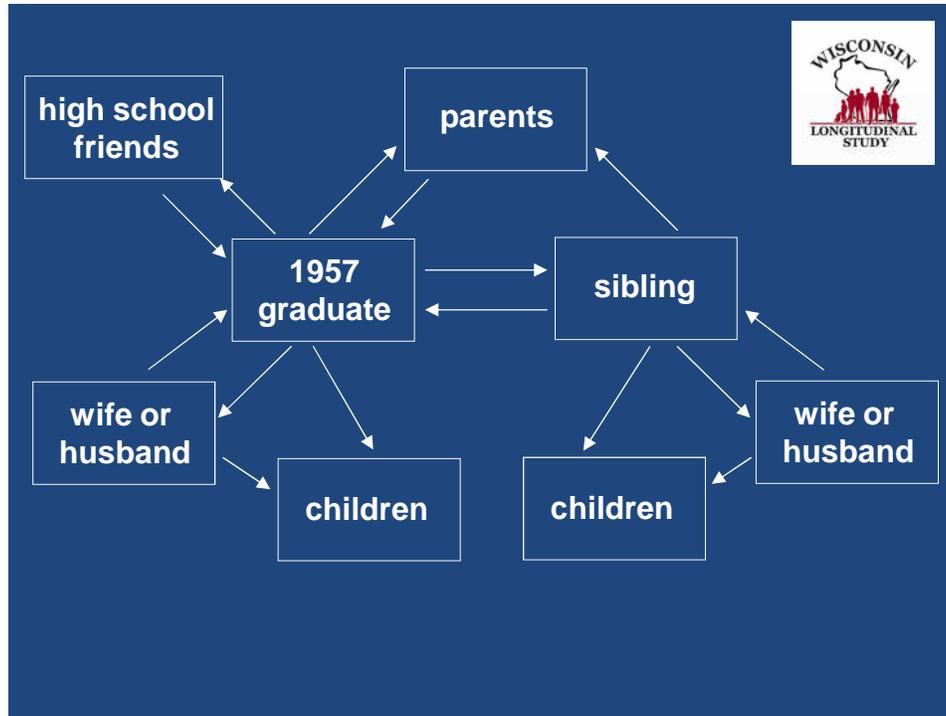
Chart 1. The Wisconsin Longitudinal Study Surveys

<p>1957 - Sample of 10,317</p> <p>1964 - Short mail survey of parents, 87% response rate</p> <p>1975 - Telephone, 89% response rate</p> <p>1977 - 2000 randomly selected siblings, telephone 80% response rate</p> <p>1992 - 4 telephone/mail; 85%/80% response rates For graduates/siblings</p> <p>2004 - 6 telephone/mail; 86%/89% response rate</p> <p>2010 - In- person and mail</p>

The WLS also collects data it labels "multidisciplinary." These data are constructed by psychologists and faculty in the medical school at the University of Wisconsin, Madison. It includes DNA collection with orogene saliva kits from graduates (and their siblings). The WLS has DNA from over 4,500 sample participants and 2,500 from their siblings, It also has 95 single nucleotide polymorphisms (SNPs) for such traits as breast cancer, cognition, depression, diabetes, impulsivity, fertility, longevity, obesity and other illnesses. The 2010 WLS is collecting health measures for physical

functioning (grip strength, timed gait test, chair rise, and peak flow measure), anthropometric measures (height, weight, waist and hip circumference, and photograph), vision screenings, health literacy, and repeat measures of prior cognitive tests. These data are used by sociologists, but have become widely used by other disciplines.

Diagram 1. Relational Structure of the Data in the Wisconsin Longitudinal Study



Data Accessibility

The WLS is committed to easy public availability of all permissible data. It has many tools on the web to facilitate use of WLS data and access to publications. All data, codebooks, questionnaires, and general documentation (i.e. scale construction) are online. There are multiple mechanisms online that can be used to search variables across survey years. There, however, are challenges when dealing with private data. That is, much of the data such as DNA samples and social security data are restricted data. How do we make it possible to grant people access to use the data is a question with which we grapple. User outreach is conducted at data workshops at multiple professional meetings, and through a small grants program. The WLS plans to target graduate students and young scholars more effectively.

Future Plans for Promotion and Dissemination

The WLS is working on a website redesign and the new site will generate easy-to-use longitudinal WLS files. It is also working on ways to ease the administrative challenges of private data access without compromising privacy issues. The WLS will make available teaching modules so that faculty analyzing the data can use these data in graduate and undergraduate courses. The challenge is how to capture new audiences outside of the existing user profile (sociology) and figure out how to target a range of disciplines and make them aware of the data. This, however, becomes problematic when potential users are located at universities who do not have access or ways for protecting the privacy of the data. This limits access at some universities.

David Howell

University of Michigan

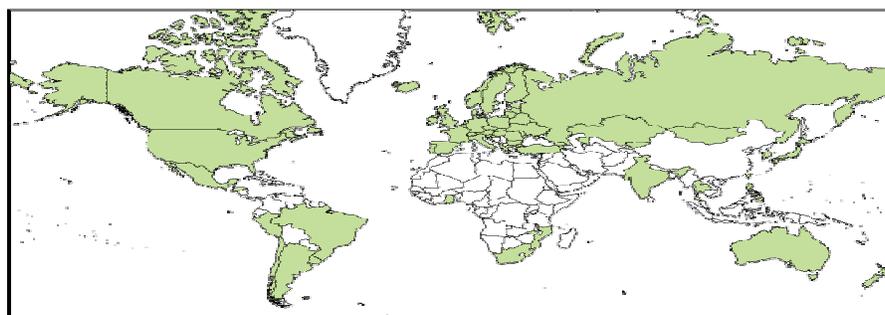
Assistant Director, *Center for Political Studies (CPS)* &

Director of Studies, *Comparative Study of Electoral Systems (CSES)*

The Comparative Study of Electoral Systems (CSES) project was founded in 1994 to promote international collaboration among national election studies. CSES is a collaborative program of research among election study teams from around the world. Participating countries include a common module of survey questions in their post-election studies. The resulting data are deposited along with voting, demographic, district and macro variables. The studies are then merged into a single, free, public data set for use in comparative study and cross-level analysis.¹³

The CSES module is a 10-15 minute respondent questionnaire with a specific substantive theme. A new theme and questionnaire is developed every five years. The data from all countries are merged into a single data set along with administrative, demographic, district, and macro variables. The CSES has a micro-macro design. The public opinion survey questions are included in each participant country's post-election study. "Micro" level data include vote choice, candidate and party evaluations, current and retrospective economic evaluations, evaluation of the electoral system itself and standardized socio-demographic measures. District level data on electoral returns, turnout, and the number of candidates are included for each respondent. Also, a system or "macro" level data report provides electoral returns, electoral rules and formulas, and regime characteristics. This design allows researchers to conduct cross-level, as well as cross-national analyses, addressing the effects of electoral institutions on citizens' attitudes and behavior, the presence and nature of social and political cleavages, and the evaluation of democratic institutions across different political regimes.

Module 3. Collaborators 2006-2011



 **CSES**
COMPARATIVE STUDY OF ELECTORAL SYSTEMS

Data Availability and Distribution

CES provides free public access to data without embargo and restrictions. Data can be downloaded at www.cses.org; and it is also archived at the Inter-university Consortium for Political and Social Research (ICPSR), GESIS - Leibniz Institute for the Social Sciences in Germany, and elsewhere. There is a standard data center to download a zip file with data and codebook and source data. The users have access to online help files that provide instructions on how to use the data sets. An online analysis tool is facilitated through the company JD Systems (JDS) in Spain.

¹³ Description from CSES webpage- <http://www.cses.org/>.

Possible Directions for the Large-Scale Surveys

Data dissemination, broadly defined, is not just about the files but also outreach, community building and maximizing scholarly output. Data dissemination requires resources. We must first ask ourselves if this is something we are uniquely qualified or positioned to provide? Are there opportunity costs? Are there partners available, including ones outside of our industry?

Where to go from here? We must first have common, machine-readable formats. These would allow customized and dynamic documentation, facilitate long-term archiving, have the ability to translate well into future technologies, and import data and documentation into third-party applications. The archiving properties should be broadly accessible and minimize reliance on proprietary storage formats. Second, we need more collaboration with third parties (sometimes we look too internally). This includes collaborative training, tools, archives, and supplemental data that allow easy “hooking into” our products. The expansion and impact of dissemination can be enhanced through training that includes better technical and analysis documentation online, expanding the potential users of the data sets, and customized resources for students. We also could better leverage technology through interactive, dynamic online experiences and multimedia platforms. We could provide more access for the non-technicians and general public through data books, pre-packaged results and reports. There is a need to do outreach for better access by high schools and non-technicians with tools that allow people to explore the data without having to download it.

The user community can be expanded by assigning more resources to community building. For example, the opportunity exists for connecting and informing communities through technology, such as through blogs, forums, and listservs. Data use could be increased by leveraging key moments, for instance, the elections year for ANES. Finally, it would be informative to complete a usability analysis to answer questions about what do people really need, and where they have problems. There are methodologies we can borrow and likely existing useful data that could be used to complete this analysis.

As one moves forward with a focus on greater data access and dissemination there are things about which to be cautious. Survey projects as application developers run the risk of becoming over-diversified, having high ongoing costs and bad economies of scale, creating prioritization challenges for small staff, becoming over-committed to a technology, and forcing specialized training on users. The possible exception is when surveys have very complicated designs, there is a need to have specialized training, but others are likely already doing this. In such instances, collaboration should be tried or an independent project pursued. A second challenge is not to be all things to all people. Distributing too many proprietary file formats could result in high quality control and user support costs, and have the unintended consequences of credibility problems such as version control, consistency, and replication problems. Resources need to be allocated to where they can best serve. Finally, too much centralization can create problems. The goal should not be for everyone to come to a particular survey project website since this could discourage innovation. People seek out different experiences and many people download the data elsewhere. Thus, it is important to have third-party redistribution policies and to make sure users are pointed back to the project for updated information. Care should also be taken that distribution practices do not delay releases since some users want ease while others want speed.

Andrew Beveridge
Queens College and CUNY,
Professor of Sociology & Developer of
Social Explorer

Social Explorer was developed in 1999 as a simple data access "online" research tool to provide easy access to historical census data and demographic information about the United States from 1790 to 2000. The easy-to-use web interface lets users create fast, intuitive, and illustrative maps and reports to visually analyze and understand demography and social change throughout history. Through a relationship with the *New York Times*, we found maps from 1910. We had all of the aggregate data from 1790 and imputations of change back to the census tract level for the American Community Survey (ACS). We digitized these and then put them on the web. This was followed by putting the New York census on the web, then the Los Angeles census, and then creating the National Science Digital Library.

The current Social Explorer includes data from the entire United States census from 1790 to 2000, all annual updates from the American Community Survey up to 2008, original census tract-level estimates for 2006 and 2007, the Religious Congregations and Membership Study from 1980 to 2000, and 2002 Carbon Emissions Data from the Vulcan Project. The site is updated continuously with new data and features. The 2010 US Census will be available shortly after its release in 2011. A full and complete reference of historical and modern census data, Social Explorer is an end-to-end solution that meets the needs of researchers and scholars without sacrificing ease of use for non-experts.

History of Social Explorer

Social Explorer was first conceived in 1999 with the mission to build the most informative and easiest to use demographics website in the world. In order to present a clean and simple interface for all users to display and extract information, Social Explorer undertakes massive data operations—5 trillion CPU operations for one census alone. The site first launched in 2003 and became available via subscription in 2007. Users from around the world access data using Social Explorer; 2.5 million maps are generated each year.

For US census tracts, the Social Explorer interface goes back to 1790 to make it easy to get to the data. There are free versions of Social Explorer and a student edition of Social Explorer by Pearson Publishing. The project, however, waived the intellectual property rights and set up as a private company to sustain the research. . Oxford University Press partnered with Social Explorer in 2010 to enable it to expand and enhance the features, functionality, and data, and reach a larger audience. A premium version is licensed to the libraries; the distribution will be done by Oxford University Press. At the recent American Library Association meeting, 200 libraries asked for trials. Currently, there are 150,000 individual users. So, there is a great deal of interest in the data, if interested persons can gain access.

Social Explorer can be illustrated by simply typing in an address. The user will receive aggregate data, along with access to where the data are located, information on what data were combined as well as access to the codebook and the questionnaire. More and more researchers are using the Social Explorer. The data can be used to develop maps to track certain topics. For example, it is possible to develop a map of the location of all religious congregations in the United States or to have a blog that track vampires. A pretty good knowledge of what users want comes from the project's experience with the user community and journalists. The Social Explorer project is moving into developing curricula; currently there is one focused on poverty and another on residential segregation.

A Social Explorer type system and interface for survey data would advance survey data access and dissemination. The system must be driven by meta-data to meet current web standards. The Social Explorer interface is a little old, but the project enjoys popularity because the interface is a very good one,

meeting user needs. When Social Explorer was developed, usability testing was not done. This would not be the case now, since there more sticky users. As aficionados of survey data and the use of maps, and having created a data dissemination interface, please note that the users are not necessarily so.

Nirmala Kannankutty

Senior Advisor/Senior Social Scientist

National Center for Science and Engineering Statistics, National Science Foundation

Data Collection and Dissemination on the S&T Enterprise:

Activities of the National Center for Science and Engineering Statistics (NCSES)

The *National Center for Science and Engineering Statistics* is a federal statistical unit within the National Science Foundation (NSF). As one of the divisions in the Directorate for Social, Behavioral and Economic Sciences, it has a clearinghouse mandate to provide data on the health and status of the U.S. science and technology (S&T) enterprise. The NCSES conducts surveys, synthesizes data, promotes gathering of comparable international data, and disseminates information via reports, electronic products, databases, and data files. The NCSES serves a broad range of customers that include policymakers, academic decision makers, academic researchers, nonprofit organizations, professional associations, the media and the general public. Its data collection efforts focus on two major areas--(1) human resources and (2) research and development (R&D).

Human Resources Data

The NCSES collects data on human resources at major stages on a continuum--precollege education, undergraduate enrollments, undergraduate degrees, graduate enrollments, graduate degrees and the workforce. The overall goal of the human resources data collection is to determine the number of people at key stages, their demographic characteristics, short-and long-term trends, as well as how performance at one stage relates to subsequent stages. The NCSES human resources data are used to develop a comprehensive picture of the S&T enterprise and comprise a series of data collections of both individuals and institutions.

1. **Surveys of Individuals.** The Survey of Earned Doctorates (SED), SESTAT (Scientists and Engineers Statistical Data System) Survey of Doctorate Recipients (SDR), National Survey of College Graduates (NSCG) and National Survey of Recent College Graduates (RCG)
2. **Surveys of Institutions.** Survey of Graduate Students and Post Doctorates in Science and Engineering (GSS).

Research and Development (R&D) Data

The NCSES studies the research and development (R&D) system of the US. It examines all major aspects of the system, including funders, performers, infrastructure, expenditures, and to outputs. The NCSES collects information to describe each component and the relationships among components, to determine long-term and short-term trends and to better understand the system as a whole. The NCSES also conducts a series of surveys of institutions and organizations annually to collect data on R&D.

Chart 1. Research and Development Surveys

Higher Education Research and Development Survey (HERD) <i>formerly known as Survey of Research and Development Expenditures at Universities and Colleges</i>
Survey of Federal Funds for Research and Development (Federal Funds)
Survey of Federal Science and Engineering Support to Universities, Colleges and Nonprofit Institutions (Federal Support)
Survey of Science and Engineering Research Facilities (Facilities)
Business R&D and Innovation Survey (BRDIS) - <i>formerly known as the Survey of Industrial Research and Development</i>
Survey of State Government R&D (State R&D)

Data Access and Dissemination

NCSES data are disseminated through a number of vehicles that include info briefs, detailed statistical tables, periodic overview reports, topical reports and working papers. Info briefs are 2-3 page highlights of results from recent surveys and analyses. Detailed Statistical Tables (DSTs) contain extensive tables from a particular survey (in electronic form only). The periodic overview reports consist of *Science and Engineering Indicators*, *National Patterns of R&D Resources*, and *Women, Minorities, and Persons with Disabilities in Science and Engineering*.

Data can be accessed online using several data tools and facilities. WebCASPAR compiles databases with a tabular data generator that provides easy access to a large body of data collected from U.S. academic institutions on topics such as degrees awarded, enrollments and R&D expenditures. It is a useful tool for computer-aided science policy research. The Scientists and Engineers Statistical Data System (SESTAT) Data Tool is an online table generator for NCSES's three surveys of the science and engineering workforce. SESTAT allows individuals to come in and create their own custom tabulations. The NCSES had worked on SESTAT for 15 years and the resultant system captures comprehensive metadata and has the facility to search for variables and full provenance. All SESTAT standards for the metadata go back to the origination of question creation and it has a very strong paradata data collection system. The two concerns of SESTAT are whether people will use the data, and having made the data so easy to use, will users ignore the origins of the data.

The NCSES will pilot a Secure Data Access Facility (SDAF) at the National Opinion Research Center (NORC) Data Enclave in the near future. The plan is to build a public access portal for confidential microdata data from NCSES surveys. In the pilot a table generator will provide access and statistical disclosure limitation procedures to the Survey of Earned Doctorates (SED). This is necessary because there may be users who try to figure out the numbers of suppressed data. The overarching reason for creating a secure data access facility is to make it easier to distribute microdata files and increase access to the survey of earned doctorates. The NCSES provides access to microdata through public-use files and to confidential data through restricted-use files. Microdata files with confidentiality protections applied are downloadable for some surveys and for others accessible through WebCASPER. Selected restricted-use files are available under a licensing agreement.

Meeting the needs of the wide range of NCSES data users is a challenge. So far, none of the data has been available at the Census research data centers. Future plans are to place data into other outlets used by researchers and publicize the availability of data. The latter has been very difficult, but they now publish a quarterly newsletter announcing new data releases. The NCSES does not have the technical expertise to build custom systems, but it is exploring what is available that with slight modifications could meet the needs of the user community. There is a much broader applicability of the NCSES data for social science and policy research. They look forward to communicating with that larger data user community.

Chart 2. Data Access Summary

Surveys/ Datasets	SRS Publications			Online Data Tools/Facilities			Microdata	
	Info Briefs	Detailed Statistical Tables	Other types	Web CASPAR	SESTAT (Science & Engineering Statistical Data System)	Secure Data Access Facility	Public Use Files	Restricted Use Files
Survey of Graduate Students and Post Doctorates in Science & Engineering	X	X	X	X			X	N/A: full public release
Survey of Earned Doctorates	X	X	X	X		X	No	X
Survey of Doctorate Recipients	X	X	X		X	X	X	X
National Survey of Recent College Graduates	X	X	X		X	future	X	X
National Survey of College Graduates	X	X	X		X	Only released as part of the SESTAT integrated file, except as a PUF in baseline years (1993 and 2003)		
SESTAT - Integrated	X	X	X		X	future	X	X
Higher Education Research & Development Survey	X	X	X	X			No	No
Federal Funds	X	X	X	X			No	No
Federal Support	X	X	X	X			No	No
Facilities	X	X	X	X			No	No
Business & Innovation Survey	X	X	future				No	Future access through Census RDC
State R&D	X	X	X	X			No	No

² IPEDS data files and data tools are also available through the National Center for Education Statistics at <http://nces.ed.gov/ipeds/>.

Vincent Hutchings and Simon Jackman

University of Michigan and Stanford University

Principal Investigators

American National Election Studies (ANES)

The American National Elections Studies (ANES) is a 60 year time series of innovative data collections. The ANES data collection began in 1948 at the University Michigan. Beginning with the 2008 election cycle, Stanford University became a co-director of ANES. ANES conducts national surveys of the American electorate in election years and carries out research and development work through pilot studies. The ANES has time-series election year data from 1948-2008. In presidential election years, the study is typically conducted both before and after the election (that is, a pre-election study and a post-election study), while for congressional election years the study has typically been conducted only after the election (a post-election study). *Pilot Studies*, normally conducted in years when there is not a national election are done to test new, or refine existing instrumentation and study designs. The ANES data are used by tens of thousands of researchers, journalist, students and citizens around the world. The ANES is designed to be used by those who are not experts and its major goals are to:

1. explain vote choice in presidential and congressional elections;
2. elucidate variation in voter turnout, i.e., broader political participation and the factors in turnout and participation;
3. facilitate socially relevant analyses pertinent to electoral context through making accessible its time series data; and
4. provide data (multiple variables) that can be used to evaluate a range of hypotheses.

To gather reliable data from minority groups, the ANES included African-American and Latino oversamples (along with Spanish language interviewing) in the 2008 and 2012 studies. The ANES project deemed it important to establish baseline data for future iterations of the study, especially in light of the candidacy and election of President Obama. Also, the incorporation of African-Americans and Latino views makes the data set more attractive to a broader range of researchers. The project also seeks partnerships with other investigators and projects to expand the reach, scope, and impact of the project. For example, the Cooperative Congressional Election Study (CCES) arose when the ANES stopped focusing on midterm elections. The ANES plans to partner and dialogue with CCES on instrumentation for the 2012 surveys.

The ANES is a long-term participant in the Comparative Studies of Electoral Systems (CSES). CSES is a collaborative program of research among election study teams from around the world that collect data on a common module of survey questions in their post-election studies. The association with CSES helps to make the data more accessible through its inclusion in an additional dissemination outlet.

Central Elements of the Current ANES

The ANES will conduct surveys in the 2012 pre-election and post election cycle to collect data for the time-series study. It will complete face-to-face interviews with 2000 citizens. It will use multi-stage cluster sampling. The sample will be a self-representing sample plus oversamples of African-American and Latino citizens. It will conduct companion Internet studies for the 2012 pre/post-election study using instrumentation nearly identical that those used in the face-to-face interviews. As mentioned above, the ANES recontacted and reinterviewed the 2008-2009 ANES Panel, in spring 2010 to measure electoral change from 2008 voters. In 2010-2012, the ANES will complete an *Evaluations of Government and Society Study*. This is a rolling (Internet-based) cross-section study of the electorate. Approximately 5-6 surveys will be conducted beginning in fall of 2010. Each will cover special topics

such as midterm elections, race and politics. These studies serve dual purposes in that they will pilot new instrumentation and track important political developments.

Transparency and Access

The ANES has an Online Commons to encourage participation of researchers and the broader user community. The Online Common is an open-call process for the user community to propose new instrumentation. This process encourages researchers and general users to advocate for new questions or suggest questions that should be continued as well as provide comment on other posted proposals. This makes the survey development more transparent in that it continually engages the online user community. Trends in usage of the online commons suggest that people are getting accustomed to and like this format.

The ANES is fielding 5-6 cross-sectional surveys between 2010 and 2012 as part of it *Evaluation of Government and Society Study*. This study provides the opportunity to pilot a larger share of items proposed by the user community for inclusion in the 2012 study. The piloting is primarily done over the Internet and is a useful approach to gauge the importance of issues proposed. In 2010, the ANES also re-contacted participants in the 2008-2009 ANES panel study. It received 34 proposals for waves 1 and 2. Wave one was completed in July 2010, and wave two was fielded in October 2010.

To further access and participation in the ANES, in 2008 the ANES added "marginal minutes" to the study. The marginal minute allowed researchers to purchase additional question space at or near the end of the post-election module. The ANES plans to extend the marginal minutes option on the 2012 survey.

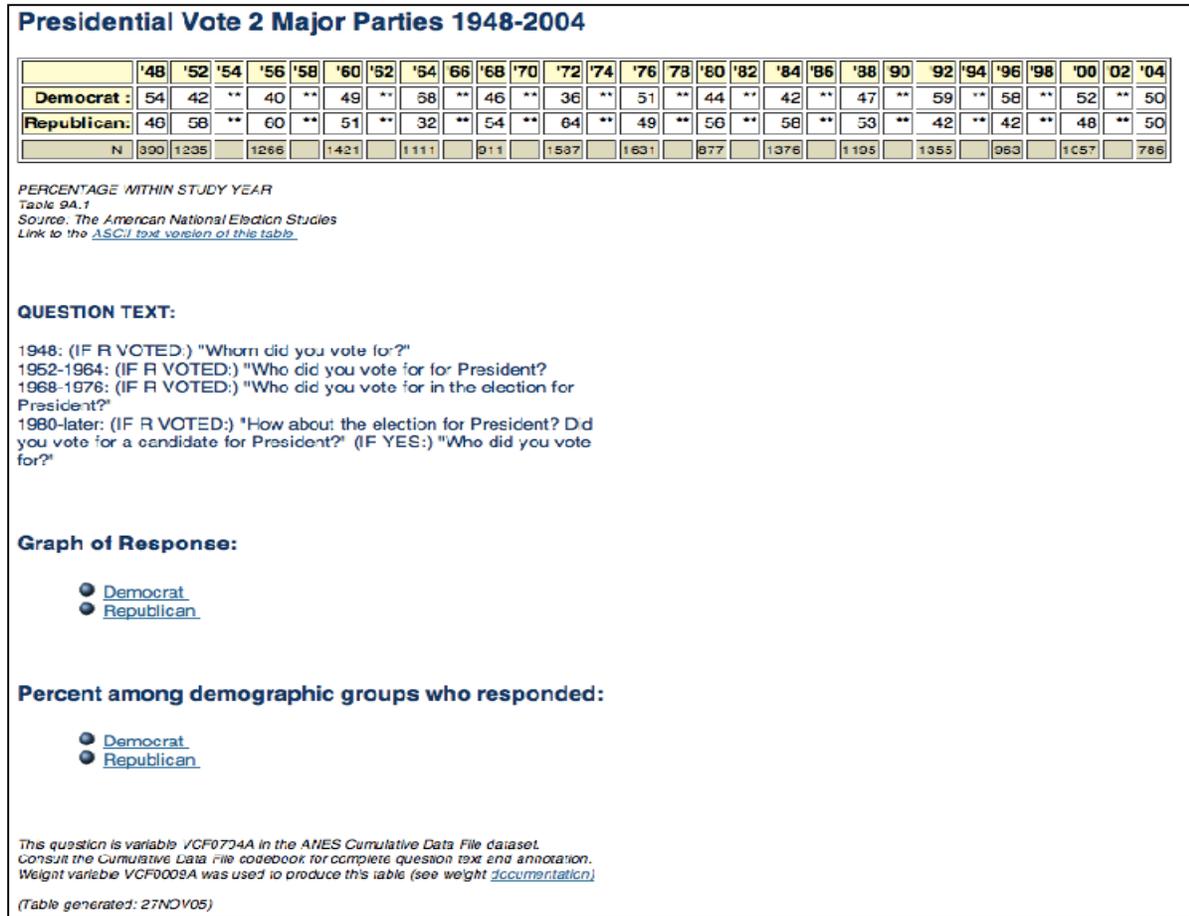
The 2008 survey consisted of pre-election fieldwork from September 9 - November 13, 2008 plus post election fieldwork from November 4 - December 30, 2010. The average face-to-face interview was 73 minutes for the former and 91 for the latter. The ANES provided an "advance" release of the interview data on March 5, 2009 and the cumulative file on June 24, 2010. The cumulative file was not entirely comprehensive, though it contained the majority of the variables from 1948 to 2008. This was a much faster release because traditionally these data are released in the year following the election.

Restricted ANES data can be accessed. In the past if a user requests census data, the request goes to the ANES principal investigators and the review board. A discussion is held and a determination made about the confidentiality protocol. The ANES has made a change to it policy whereby date the US census tract is the primary sampling unit (PSU). The project acknowledges that past decisions about data release might invite more risk of disclosure than desired. Further, the nature of the way in which the data will be collected for 2012, coupled with the merger with other data sets, could make this risk greater than desired.

It is critically important that the ANES provides free and timely access to data since it is a federally funded project. While this is important, it is not the most important issue; more challenging is the ability to provide quality data at a reasonable cost, with a very low risk of disclosure. Data dissemination is stressed as a National Science Foundation priority, but not necessarily the top of the priority list for the ANES team, who are there primarily to conduct the survey.

The primary mode of dissemination for ANES data is its website electionstudies.org. It is also deposited at ICPSR (Inter-university Consortium for Political and Social Research). The ANES files are not "long" (49,760 records from 1948-2008), but are "wide." The ANES is currently supplying data in SPSS file formats. It devotes substantial resources to the production of the cumulative data file, *Guide to Public Opinion* (see illustration below), and creating "how to analyze" guides for oversamples, split balloting and to address the surge in interest by political scientists in ANES sampling.

Figure 1. ANES Guide to Public Opinion



Expanding the Use of ANES

The ANES was cited 5,620 times in books, journal articles, and new articles where the data is referenced as of January 2010. Expanding the use of ANES requires getting beyond the "flat file" distribution method. Perhaps there are online analysis tool or "off the shelf" solutions that could be used. Some of the dissemination is being downstreamed to Survey and Data Analysis (SDA) at the University of California-Berkeley and Stanford University library DEWI (Web-based search and extraction tool). The PIs would like guidance on what needs to be machine actionable; should the focus be on "standards" versus "technology?"

ANES faces several documentation challenges in its stewardship of the long range of studies dating from 1948. Traditional codebooks are cumbersome, given study complexity (especially in recent years). For example, newer waves of data have split ballot, minority oversamples, mode experiments, and panel designs. There are different coding schemes for ostensibly the same variables across different studies; sub-group analysis, pooling over multiple years (e.g., Southern whites); and a resurgence of interest in sampling and data-quality within the context of changes in the sampling design over time.

The current practices of the ANES also face challenges. The documentation and dissemination of data are not tightly coupled with data generation (i.e., instrument design, sampling, and field operations). ANES data is collected by Internet vendors and this process produces meta- and para-data. Should these data be included as deliverables? If so, that what are the data standards? The Internet data collection sector is an interesting world with which to work, but can Internet vendors be held to the same standards?

Metrics present another challenge. It is easy to track citations to study, but to get maximum use of these there must be the capacity to "drill-down" variable-by-variable. It would be very useful to employ the National Opinion Research Center (NORC) access to data approach to track the level of access by the users. This type of tracking would allow the ANES to get information back about what variables are being used. Would DDI (Data Documentation Initiative) or another project or platform help with currently resource-intensive deliverables (e.g., updating cumulative files and *Guide to Public Opinion*). It seems that the ANES project is trying to figure out the same issues with which other projects are also wrestling, within an environment of resource constraints. Are there success stories from which we all can learn? Where or who is doing this right? The ANES, with a 40-year legacy of data and responsibility to continue to collect data, has to serve a dual function of collecting data and disseminating data. Is there a good model?

Peter Marsden and Tom Smith

Harvard University and National Opinion Research Center (NORC)

Principal Investigators, General Social Survey

Dissemination Strategies and Challenges for the General Social Survey (GSS)

The General Social Survey (GSS) has three overarching goals. These are: (1) assemble data on sociological issues, (2) place the US in context of other countries within the international social survey program; and (3) to get data to multiple users, one of the initial objects of creating the GSS. The current GSS is a repeated cross-section survey with a short-term prospective panel. The survey's target population is non-institutionalized adults who speak either English or Spanish. The average length of an interview is 90 minutes, and with the present short term prospective or longitudinal design requiring a re-interview of respondents at two- year intervals.

The GSS employs a multistage area probability sampling design for household selection. One respondent per household is interviewed and there is a subsampling (or double sampling) of nonrespondents. The weighted response rate is above 70 percent and the panel retention rate was approximately 80 percent in 2008.

Table 1. Completed and (Projected) Interviews in GSS Panel Design

Panel	Survey Year				
	2006	2008	2010	2012	2014
2006-2010	2000	1536	(1230)		
2008-2012		2023	(1550)	(1240)	
2010-2014			(2000)	(1500)	(1200)
2012-2016				(2000)	(1500)
2014-2018					(2000)

The GSS content consists of three components: a replicating "core" set of items, international modules and topical modules. The GSS replicating "core" comprise socio-demographic background items with replicated measurements of sociopolitical attitudes and behaviors. Many core items are on three overlapping ballots answered by a random two-thirds of each sample. The GSS participates in the International Social Survey Program (ISSP) and includes international modules on the GSS. Finally, topical modules that are ordinarily only asked one time are the third type of data collected on the GSS.

Dissemination

GSS data are disseminated through many major data archives and distributors. The Roper Center, University of Connecticut, Inter-Consortium for Political and Social Research (ICPSR), University of Berkeley, American Religion Data Archive, and Zentralarchiv, Cologne (for ISSP). Other dissemination vehicles for GSS data are *GSSNews*, the project's newsletter, and the GSS website. Project staff members conduct special and poster sessions at professional meetings and policy related activities. Users may get access to restricted data via arrangements within NORC who manages distribution depending on the disclosure risk.

Data are distributed via several websites. The main website at NORC (www.norc.org/GSS+website); the main ISSP site (www.issp.org); and the Survey and Analysis site at the University of California at Berkeley (sda.berkeley.edu). The NORC website contains information on the data that are available, online analysis tools and data documentation.

Table 2. GSS NORC WebSite: Data, Analysis & Documentation

www.norc.org/GSS+website	
Data	<ul style="list-style-type: none"> • Downloads in SPSS and Stata • Entire 1972-2008 cum file • Cross sectional files for 27 years • 2008 panel cross section file • 2006-2008 panel file • Customized data sets can be constructed via NESSTAR and SDA • NESSTAR supports 12 formats
Online Data Analysis	<ul style="list-style-type: none"> • NESSTAR permits crosstabulations, regression analysis and graphics • Link to SDA permits frequency distributions, crosstabulations, comparison of means, correlation and regression
Documentation	<ul style="list-style-type: none"> • <i>Cumulative Codebook</i> including detailed specifications on sampling and fieldwork • Release notes on updates versions of data sets • Questionnaires • Protocol for obtaining restricted data • Content of replicating core • Browsing of variables by mnemonic (alphabetical), sequence, collections, (topical, ISSP modules) and subject index. Also, provides tabulations, trends by year

Future Directions

The GSS data collection and dissemination will move to "interview to Internet." This will include the XML export of CAI (computer-assisted interview) data for use in Internet environment. Documentations and metadata will be transferred from CAI instruments to the dissemination portal. Plans are to use the Data Documentation Initiative (DDI) standard. This will reduce the interval between the collection of data and data availability. The goal is to make data collection and dissemination go hand in hand. The GSS plans to have expanded metadata. It will append extensive metadata on individual questions to all occurrences of a variable via hypertext links. The metadata would encompass all site documents--codebook, appendices, reports and bibliographic references. Links would also be available to appendices and user publications. Of the thousands of GSS citations, currently about 6,000 citations have been abstracted, including the full citation and the variables used. The major limitation to the linking to citations is that it is very labor intensive.

The GSS project will work in the future to expand the GSS online analysis capability. The project will continue its collaborative work with Nesstar, SDA, and other platforms. In order to meet the increasing demand for the ability to do more online graphics, like organized line graphs to show trends pictorially, the GSS will expand its graphics capabilities. Finally, the project recognizes the need and plans to expand its documentation and support. Specifically, it would like to develop:

1. Online tutorials/webinars on conducting GSS analyses
2. Additional FAQs
3. Links to textbooks that use GSS
4. Updating of bibliography to expand coverage in recent years
5. Links to forums, newsgroups - particularly for students
6. Collaborative tagging

Collaborative tagging would enable users to create tags, and consequently, their own user index that would be made publicly available. This would address the issue of how best to aid users with GSS content (e.g., find a variable of interest) in a data collection with over 5000 variables. The collaborative tagging project would rely upon the user community.

Narayan Sastry and Robert Schoeni

University of Michigan

Principal Investigators, Panel Study of Income Dynamics (PSID)

*Dissemination Tools, Services, Challenges &
Future Vision for the Panel Study of Income Dynamics (PSID)*

Background

President Lyndon Johnson's War on Poverty in 1960s stimulated interest in understanding the dynamics of income and poverty. The Panel Study of Income Dynamics (PSID) was established in 1968 as a 5-year project to study income dynamics within a national sample of 5,000 families. Today the PSID includes 9,000 families and is used to address a much broader set of issues.

Members of the 5,000 families first interviewed in 1968 are followed as they grow and establish their own economic family units. Therefore, multiple generations of family members are surveyed. To make the sample more representative of the US demographically, post -1966 immigrants (i.e., Asians and Latinos) were added in 1997. PSID study participants have been interviewed by telephone (using CATI-computer aided telephone interviewing) since 1993. They were interviewed annual from 1968-1997 and biennial since 1997. One interview is conducted in each family unit. Study core response rates range from 96-98 percent .

The PSID is unique because it is a national sample of people of all ages and the study spans the entire life course. The study is a long panel with a genealogical design. The longitudinal nature of the survey makes it possible to study a point-in-time for parent-child, grandparent-child, and siblings, and also specific life course points. The PSID context is broad and deep. For example, it collects economic (income, wealth, pensions, and consumptions), demographic (fertility, mortality, migration and marriage), health (health status, behavior and insurance) as well as child development (health, cognition and psychosocial wellbeing) data.

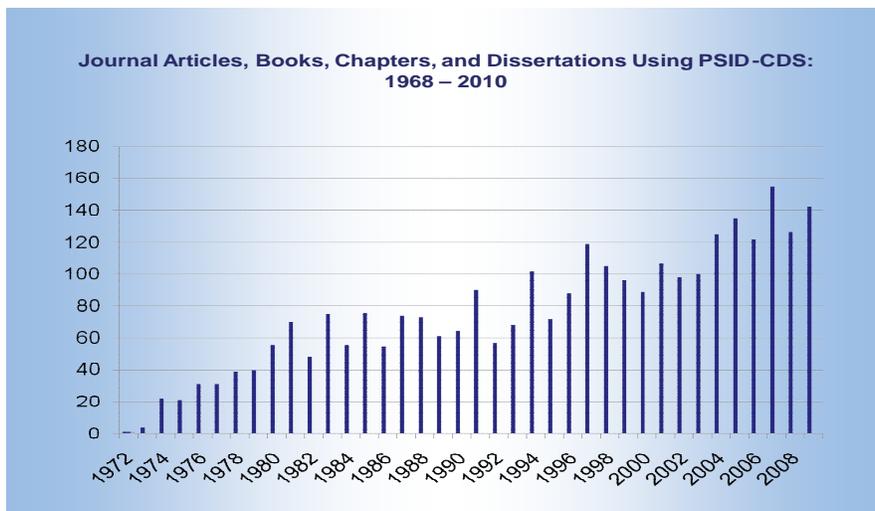
The PSID sample has the Child Development Supplement (CDS) and Transition to Adulthood (TA) studies. The CDS-TA is a nationally representative longitudinal study of children and their families that collects data on a broad array of developmental outcomes with the context of family, neighborhood, and school environments. The PSID has collected six waves on data on the children in the CDS sample and 3 from children in the TA study. The CDS-TA sample eventually became an active panel of the core PSID when they "split-off and establish their own family units. The PSID collects a wide array of data on employment and beyond. The survey questionnaire has doubled since the study's inception. The questionnaire length doubled when the conduct of the PSID was moved to every other year.

The PSID has an active user community. These data are cited in 2,944 peer reviewed publications and nearly 400 publications. There were 4.84 million hits to the PSID website in 2009, 12,800 registered data users in total, and 122 active contracts (301 sensitive data contracts in total). The PSID serves as a model for surveys in other countries and is replicated in numerous countries across the world. The PSID is a tool for the federal government. The Treasury Department, Departments of Agriculture, f Housing and Urban Development, and Health and Housing Service, and the Congressional Budget Office use the PSID data for policy research and analysis.

Chart 1. Example of Types of PSID Data

- ▶ Employment
- ▶ Computer use
- ▶ Income detail
- ▶ Expenditures
- ▶ Program participation
- ▶ Housing
- ▶ Housework
- ▶ Child care
- ▶ Marriage & fertility history
- ▶ Geographic detail
- ▶ Health conditions and behaviors
- ▶ Mental health
- ▶ Vehicle inventory & expenses
- ▶ Education & utility expenses
- ▶ Wealth & active savings
- ▶ Philanthropic giving
- ▶ Demographics & education
- ▶ Mortality and cause of death linked to NDI
- ▶ Links to Census Geocode, CCD-PSS, IPEDS, AHS

Chart 2. Panel Study of Income Dynamics (PSID) Data Usage



Possible Future PSID Innovations

The PSID is considering innovations in its interviewing protocol, sample design and size, and survey content. In the future, the PSID would like to interview all individuals age 16 and older, follow attritions into perpetuity and follow non-sample parents and step-parents. The project would also consider increasing the sample size to improve representativeness by adding an immigrant refresher sample (last done in 1997). If possible, the PSID would collect data in several new content areas. These

include information on newborns, time use and well-being among older adults, retrospective reports on adverse childhood experiences (maltreatment, social interactions, and social isolation), and expanded personality, psychosocial measures, and bio-measures.

The PSID would like to pursue new data collection methods. Specifically, a mixed methods mode would be explored using web-based methods, CATI, face-to-face interviews, and mail surveys. The project would also expand its interviewing of dependents. Another area of potential fruitfulness would be to expand administrative links to the Social Security Administration, firm data, earnings and benefits data, and Medicaid data. The PSID plans to pursue further development of geospatial data and adding a methods panel of 2,000 families as a study testbed.

Current Dissemination

The PSID data is available through the On-line Data Center (www.psidonline.org). The center provides access to public use data and supports customized subsetting of all waves. Subsetting is done by data file (i.e. family, individual, Child Development Supplement, and Transition to Adult) or by using cross-year index by topic. The Center provides customized codebooks, and facilities for variable searching and browsing, and complex data set merges. Data are available in SAS, STATA, dBase, ASCII, and Excel. The PSID releases data via packaged files. The entire PSID data set can be downloaded directly and separate files for supplemental data are available. The supplemental files have distinct data structures the event history calendar and time diaries. Early-release data on the 2009 housing, mortgage distress and wealth was recently provided. Finally, PSID restricted data are available through contractual agreement.

Chart 3. Restricted Data Available Under Contract

- ▶ Geocode Matched Data File 1968-2007
 - Identifies 2000 census tract
- ▶ Assisted Housing Data
 - Links address to HUD data on subsidized housing
- ▶ Mortality Data File 1968-2007
 - Date and cause of death from NDI
- ▶ Common Core / IPEDS Data from NCES
 - Links school characteristics to families in CDS and TA
- ▶ Hurricane Katrina Data
 - Supplement for n=555 families
 - Effects on socioeconomic and psychological well-being
 - Extensive data on substance use

The PSID has tutorials that make it easier to use and understand the data by providing a step-by-step guide to answer specific questions. It demonstrates various features, for example, cross sectional, panel and intergenerational analyses, and illustrates several research domains. All tabulations can be conducted within Excel and the tutorials are targeted to new users, including undergraduates and advanced secondary school students.

The PSID's Family Identification Mapping System (FIMS) has software that facilitates the use of PSID's genealogical data by automatically creating extracts that link identification variables for related family. The mappings can be inter-generational (e.g., parent-child or grandparent-parent-child) or intra-

generational (e.g., sibling–sibling). Users can then merge data of substantive interest to do inter- and intra-generational analysis.

The PSID has several other dissemination and user tools. These include the Cross-National Equivalent File that distributes data from the PSID as well as the British Household Panel Study, Household Income and Labour Dynamics in Australia, the Korea Labor and Income Panel Study, Swiss Household Panel, Canadian Survey of Labour and Income Dynamics, and German Socio-Economic Panel. The PSID-HRS cross-walk is also available to users.

Future Dissemination Plans

The PSID project currently completes outreach to professional meetings and provides user support. New-user workshops are held at professional meetings and at the ICPSR summer institutes, and early results conferences are offered. The PSID has a helpdesk staffed by a full-time staff member. In the future the PSID would like to provide online data analysis using NESSTAR or SDA, but the data complexity and restricted data present a challenge. It would like to use DSDR/ICPSR to process restricted data applications; DSDR will facilitate the applications for restricted data. The PSID will also participate in the DSDR/ICPSR "cloud computing" initiative. There are plans to recode and rename data across years to enhance comparability, which is an enormous undertaking; restructure both individual- and household-level data files; and scan paper records for new variables and measures.

Challenges and Issues

The PSID has challenges and issues as it moves into the future. First, it will strive to make it easier for unsophisticated users to analyze the data. Second, the PSID must find ways to increase access to restricted data while maintaining security. Lastly, it has to continue to address the issue of obtaining IRB (or other) guidance for disseminating restricted data. The end users' institution IRB has to approve use of PSID, but leaving data security to the home institution IRB can become difficult.

Mark Chaves

Professor of Sociology, Duke University, and
Chair of the GSS Board of Overseers
ANES, GSS and PSID Data Dissemination

In order to fund NSF-supported survey data dissemination, there are six important questions that need to be answered if one starts with the assumption that nothing is broken. The three projects --General Social Survey, American National Election Studies and Panel Study of Income Dynamics -- are doing a tremendous job in both data collection and data dissemination. The goal is to discover how to make this better.

1. What should be disseminated? In discussing dissemination, the key issue is should dissemination focus on the data itself or the dissemination of knowledge from the data. There is some tension between the two and what activities would be funded would vary based on the dissemination priority.
2. Dissemination to whom? Is it professional researchers, students and everyone else (i.e., the media and general public)? Dissemination of data targets students and researchers, the "everybody else" group may prefer knowledge rather than the data.
3. What is the goal? What is to be accomplished by the dissemination? There are three major goal of data dissemination. First, dissemination should enhance accessibility for researchers. Second, it seeks to provide training and community building to expand the user community. Third, dissemination should enhance the visibility of these data in the public eye. This, in itself, would be a major accomplishment.
4. Is the goal to enhance dissemination of the three large data sets or to enhance dissemination for social science data sets more generally? It seems that by doing the former, it would help the latter also.
5. What should be done to improve dissemination? It is not clear whether the projects themselves are interested in facilitating broader dissemination.

With these questions a workable process might be to develop lists of what needs to be done to improve dissemination relative to the goals in question three above. Each goal involves different factors.

6. Once a list of activities has been compiled and been prioritized, the question is who should do this? Should it be the PIs of the large data sets or should this dissemination be done by others?

Once the earlier questions are answered, it is easier to address the sixth question. If the top priority is get all of the data into an accessible, usable form, the solution would be project specific. If the top priority is to create a variable level search capacity, then it would make sense to give resources to someone else to facilitate the connections. If the goal was to reach the public, and impress the importance of what we discover, then people should be encouraged to write publications accessible to the general public because journal articles are not the locations which the public accesses. Part of the reason that people do not know as much about the large data sets is because the standard model is the academic article. The theme of these comments is to clarify the distinctions between types of dissemination. This can help us figure out how to prioritize user and data needs and resources.

Darrell Donakowski and Matthew DeBell
University of Michigan and Stanford University
American National Election Studies (ANES)
Accessibility and Usability of ANES Product and Services

The American National Election Studies (ANES) user community is comprised of social scientists (particularly political scientists), instructors and teachers, students (especially graduate students), journalists, and persons in the general public with an interest in public opinion and behavior related to national politics. Data are available on the website, www.electionstudies.org, and data archives at the Inter-university Consortium for Political and Social Research (ICPSR). A version of the cumulative file is available at the Association of Religion Data Archives (ARDA) and at the University of California Data-Survey Documentation and Analysis. Since there are multiple access points, it is difficult to get a census of who is using the data.

There are two kinds of users--data analysts and consumers of generated tables and statistics, and thus two kinds of data needs. Analysts' need microdata (or an online analysis tool) and documentation. The project provides documentation to help users find relevant questions and data sets, directs them to use instructions, and it provides methodology and data quality assessment, for example, AAPOR (American Association of Public Opinion Research) reporting. The services provided to analysts and consumers vary. For example, analysts tend to need personalized help from an ANES expert. The table and statistic consumers need the ability to find relevant questions and thus access to questionnaires and content description. They also want to find or generate relevant estimates via tables and figures, and consequently want access to an online analysis tool. Would the table and statistic consumers become data analysts if they were given the right tools? If so, should those tools come from ANES or should it come from another source?

Accessibility and Usability

The ANES serves the needs of users in several major domains. These include data formats, personalized help, technical guidance, timeliness of data release, and the ANES guide to public opinion data and web usage. The format that the ANES uses to release data is text flat files, fixed width and .csf. It provides file-build code for Stata, SPSS, and SAS statistical packages and for a SPSS portable file (.por). The ANES is considering the release of full data sets in most used formats, but realizes it should be conscious of possibly overreaching in trying to commit to too many data formats. Other considerations include using the Integrated Public Use Microdata Series (IPUMS) mode of selecting variables and desired data. The release of full data sets in multiple formats requires addressing issues of quality assurance and engaging emerging standards.

Technical Guidance

The ANES offers users personalized help and technical guidance. It provides access to authentic data, for example, which provides information on which studies have what questions and variables; and assists with downloading and reading data files and variables. It assists data users with methodological issues related to the study design, sample and weighting issues. The ANES team, however, does not answer questions about other surveys, statistical and software, and population data, or provide consulting. The ANES team responds to an average of one question per day.

The ANES is expanding its methodological documentation and technical guidance. All methodological activities are now being recorded so in the event that the studies need to be replicated, this could be easily done. The ANES project is structured so that it out exists any particular project directors. Thus, comprehensive project documentation is available to access and evaluate the methodology. The ANES team provides assistance with complex sample data and analysis instructions. Many users look for

articles reporting weights, standard errors and statistical significance. The ANES now releases the documentation that explains the weighting process and provides weight selection instructions on multi-wave studies. The 2008-2009 panel study documentation includes 70 weights that analysts can use to select the appropriate weight for their particular analyses.

Timeliness of Data Releases

The ANES project does an advance release of study data. The release process is relatively fast and includes survey data only; it does not include administrative data and other variables. For instance, the 2008 time series data was released in March 2009, and the 2008-2009 Panel Study data in January 2009. The full data release is much slower. These data include additional variables, is cleaned and as error-free as the team can make it and takes 3-6 months longer to release.

ANES Guide to Public Opinion

The ANES Guides provides tables and figures with key estimates from select "core" variables. The Guide is updated with each time-series release. The Guide is well-used with 100s of downloads each month. The number of successful webpage requests for the time-series data, panel data downloads and cumulative file downloads are also steadily trending up. For example, in April there were 500,000 successful webpage requests.

Other Projects

The ANES is continuing to work on the cumulative data file and to address issues related to restricted data access. Currently it is planning a web redesign to make updating the site easier and making it generally more user-friendly. Finally, it plans to investigate concordance i.e., adding up all of the data and releasing data in additional widely-used data formats.

Collaborative Efforts

The Institute for Survey Research (ISR) has a Data Stewardship Project whose goal is to bring together all the data projects at the University of Michigan. The Project will determine common needs with plans to develop common solutions, and best practices and principles for documentation and metadata. The ANES and Panel Study of Income Dynamics (PSID) are represented on the committee.

The ANES is also working with the Roper Center and iPOLL and Survey Documentation and Analysis (SDA) to enhance dissemination and analysis. Its outreach to the user community is enhanced through the ANES Online Commons (where users can suggest survey questions and raise data and survey issues via the Internet), data archives and users groups, for example, the IASSIST (International Association for Social Science Information Services and Technology) data librarian group. The data archives are better able to communicate to users how to get the data and updates on data releases.

Michael Hout
University of California, Berkeley
Principal Investigator, General Social Survey (GSS)

New Users' Needs as the GSS Advances

Expanded Needs for Weights

The General Social Survey (GSS) continues to address new user needs, in particular, how recent changes at the GSS have required changes in access by the user community. Ongoing innovations in the GSS require new ways to serve users. The changes in the GSS expand the need for weights. The GSS panel design that began in 2006 has introduced several complexities for data analysis. Specifically, the panel design requires the ability to link panel participants overtime (obviously); and weight cross-section cases to include panel participant (or to only use respondents' first interview). The data set now contains multiple weights to address oversamples, randomization problems, household size correction, non response call back, and panel participants at re-interview and in combination with other weights.

Chart 1. Weights in the General Social Survey (GSS) Data set

- Black oversamples of 1982 & 1987 (OVERSAMP)
- Fix randomization problem, 1978–1985 (FORMWT)
- Household size correction (ADULTS)

- Non-response call-back (WTSSNR & WTSSALL)
- Panel members at re-interview & in combination with other weights (WTPAN, WTPAN, WTCOMB & WTCOMB)

Given the complexity of weighting of data, how can the GSS and others provide guidance for data analysis? The GSS codebook is the definitive source of information. It is the optimal reference for sophisticated users, but daunting for undergraduate students, novices (intermediate skill users) and journalists. Survey Documentation and Analysis (SDA) applies default weight that can be overridden or cancelled. Different statistical software also offers sophisticated ways to handle design complexities. Finally, there are textbooks to guide users, for example, (Treiman 2009) *Quantitative Data Analysis*¹⁴ and Heeringa, Brady & Berglund (2010) *Applied Survey Data Analysis*.¹⁵

Panel Data

The GSS began a panel design in 2006 (see chart 2). The panel design presents a set of issues. It is wrong to just "toss" re-interviews (for the panel) in with first interviews when doing analysis. Some users drop re-interviewed cases, but this may be an unduly conservative approach. In a forthcoming *Methodological Report*, the GSS will provide better weights. A special edition of *Survey Methods and Research* that will address issues in GSS weighting is planned for early 2012. The third wave of the first panel was completed in 2010. Generating estimates of reliabilities for all items in the three-wave panel will be a huge boost for usability of the panel data.

¹⁴ Treiman, Donald J., *Quantitative Data Analysis, Doing Social Research to Test Ideas*.

¹⁵ Heeringa, Steven G., Brady T. West & Patricia A. Berglund, *Applied Data Analysis*, CRC Press.

Chart 2. General Social Survey Panel Design

Panel	Survey Year				
	2006	2008	2010	2012	2014
2006-2010	2000	1536	(1230)		
2008-2012		2023	(1550)	(1240)	
2010-2014			(2000)	(1500)	(1200)
2012-2016				(2000)	(1500)
2014-2018					(2000)

Macro Data

Work is underway to add US census and ACS (American Community Survey) data to the GSS. Hierarchical linear models (HLM) methods are well-known to many users so these are in the current toolkit for academic users. Other users probably cannot make full use of metadata. Many users are already accessing geo-details through easy-to-get agreements. The GSS data use agreement has users incur responsibilities for data disclosure, but the GSS still risks disclosure. Macro data release would eliminate site license agreements, and it eliminates that risk of disclosure. Macro-data, however, would also limit the options of the most sophisticated users. The users sign an agreement, but the GSS is at risk for allowing the universities to grant access to the data. GSS will have to trust the universities since the question of who controls the data remains open.

Future Products

The GSS plans to do multiple imputations for missing data. Better trend estimates could be generated if the data were corrected for non-response. Also, the response "don't know" is substantively interesting to many researchers. If it would be helpful to the user community, the GSS could provide two imputations: one that corrected for nonresponses and "don't knows" and one that does not. The idea of using a four project concordance or some other option that makes the presentations of the large data sets more standard is worth pursuing. One of the reasons why people go to SDA is that the three data sets look the same. Users would probably use more than one of the three data sets (GSS, Panel Study of Income Dynamics, and American National Election Studies) if they only had to learn one interface and it works for all three survey. Or in thinking ever more broadly, an integrated portal shared by all surveys represented at the meeting would increase access and dissemination of all.

The GSS is interested, as is the case with the other data projects, in reaching out and doing the preparation to extend accessibility to less sophisticated users. It would be useful to have a meeting with this community to gauge what they need. For example, what would someone from the *New York Times* need in order to more easily access and use GSS data?

Dean Lillard
 Cornell University
 Senior Research Associate and Co-Director and Project Manager
 Cross-National Equivalent File Study (CNEF)

Enhancing Data Dissemination: The Cross-National Equivalent File (CNEF)

What is the Cross-National Equivalent File (CNEF)?

The CNEF has a different user group than the projects that are data collectors. It is the repackaging of seven panel studies from around the world. It started with a project comparing outcomes from Germany and USA panel studies. It currently includes the Panel Study of Income Dynamics (PSID) in the USA and panel studies conducted in Germany, Great Britain, Canada, Australia, Switzerland and Korea (see chart below). The CNEF provides standardized and harmonized measures of income, demographics, employment, health, and satisfaction (with life), and data sets with common variable names, concepts and common response categories. This facilitates cross-national comparative research through making analysis of separate data sets a less daunting task, provides a supporting link to underlying national micro-data, and promotes worldwide availability of these panel data for scientific research. The PSID is the "grandfather" of all of the other panel studies being done in other countries. Panel studies have become a world-wide phenomenon.

Chart 1. Cross-National Equivalent File (CNEF) Data sets



- Standardized / Harmonized Measures: Income, Demographics, Employment, Health, Satisfaction
- Worldwide availability for scientific research
- Supporting link to underlying national microdata
- C-7 (2009)
- USA PSID – Panel Study of Income Dynamics (1970-2007)
- Germany SOEP – German Socio-Economic Panel Study (1983-2008)
- Great Britain BHPS – British Household Panel Survey (1991-2007)
[USoc] – Understanding Society (coming)
- Canada SLID – Survey of Labour and Income Dynamics (1992-2006)
- Australia HILDA – Household, Income and Labour Dynamics in Australia (2001-2008)
- Switzerland SHP – Swiss Household Panel (1999-2008)
- Korea KLIPS – Korea Labor and Income Panel Survey (1999-2007)

• Frick, J.R., Jenkins, S. P., Lillard, D. R., Lipps, O. and Wooden, M. (2007): The Cross-National Equivalent File (CNEF) and its Member Country Household Panel Studies. Schmoller's Jahrbuch - Journal of Applied Social Science Studies. 127 (4): 627-654.

Why Harmonize?

There is an intense interest in cross-national comparisons. A *GoogleScholar* search on cross-national or cross national yielded 1,740,000 hits, but panel data are lacking. If “longitudinal” and “panel” are added, the “hits” fall to 151,000. Also, about one-third of the scholars actually use data. Meanwhile, comparative research has several scientific and social benefits. Comparisons of social and economic outcomes across nations can enhance jurisdictional competition and identify best policy practices. Cross-national data enables studies within different social and economic contexts that contribute to our fundamental understanding of human behavior. Research can also use these data as representative of the “world” to conduct quasi “natural” experiments because they provide greater policy variation, a much different policy mix, and an opportunity to test “out of sample”

Conceptual Basis for Harmonization

Harmonization must begin with a well-defined theoretical concept that is independent of culture and country-specific factors. The definition of concepts comes from discovering how the variables are being used and which ones are being used. Thus, harmonization of a variable is a response to the demand from the users. There are two general classes of theoretical concepts: concepts linked to the objective physical world and abstract concepts with a well-defined empirical counterpart. Theoretical concepts include age, sex, time, pregnancy, geographic location, and distance; examples of abstract concepts are tax, income, price, marital status, occupation, and industry. Conceptual variables must be measured. This is not a major problem since there are household panel studies in 30-plus countries that provide plenty of potential for harmonization. Conceptual variables, however, must also match the empirical measure within *and* across countries. This requires harmonization survey-by-survey.

The CNEF harmonization evolves in two ways: the addition of data from other countries and through expansion of the variable set. It plans to add data from Russia (Russian Longitudinal Monitoring Survey, RLMS) in 2011 and Israel (Israel Longitudinal Survey of Families, ILSF) in 2012. Other potential candidates are Mexico (Mexican Family Life Survey MxFLS) and the South Africa (National Income Dynamics Study, NIDS) in 2011. The CNEF expanded the variable set recently by adding satisfaction with life and is currently looking to add life-time smoking behavior. The philosophy that guides CNEF is that “research guides harmonization.” That is, substantive questions, theory and the end use of harmonized data, not harmonized for sake of harmonization, guides the process. It provides access to algorithms, data, and code, involves the research community checks and vets, and holds open debate on best practices and peer review.

New variables are created and added to CNEF by the user community. Potential data users develop a workplan that identify a behavior of interest. These variables are ones that are determined by biology and cultures or ones influenced by public policy. The core explanatory variables are extracted from CNEF who confirm s that the parent surveys include data on the main outcome and/or explanatory variable, and then equalize the variable or variables. The new variables are published and added to CNEF. When the research based upon CNEF data has passed peer review, the authors send CNEF the citation to the published papers that develops and uses the methods; and the algorithm (and code) used to create new variables. The algorithm and paper is cited in CNEF documentation.

Chart 2. Structure of Data Harmonization



USA	PSID	–	Panel Study of Income Dynamics
Germany	SOEP	–	German Socio-Economic Panel Study
Great Britain	BHPS	–	British Household Panel Survey
Canada	SLID	–	Survey of Labour and Income Dynamics
Australia	HILDA	–	Household, Income and Labour Dynamics in Australia
Switzerland	SHP	–	Swiss Household Panel (1999-2008)
Korea	KLIPS	–	Korea Labor and Income Panel Survey

The CNEF partners are discussing a framework to facilitate comparative research projects. Under consideration are processes or mechanisms by which questions could be added to existing studies and to collect missing data. The substance of the questions and data would be researcher-initiated and ideally policy relevant. This would move these data sets forward to become more comparable. Another current development is to consider proposals to coordinate and adapt survey questions to harmonize ex-ante. Creating theme projects from each of the contributing countries about a policy question that cuts across many countries, for example, like a retirement study done by **Jon Gruber and David Wise**¹⁶. Currently, also there is a call for proposals in Europe to fund research looking at effects of the environment on behavior.

Challenges and Potential

The data dissemination process is complicated by national policies in some countries and by the limited scope for the Internet interface. Funding is needed to keep CNEF non-research core activities in operation, to foster collaboration between existing and potential partners, and to attract and involve new researchers. Harmonization, nonetheless, offers great promise for leveraging the benefits of the panel studies to understand basic human behavior and the effects of policies, and for conducting varied "natural" experiments. It requires careful data constructive, but can add significant analytic value and fosters cooperation and exchange of ideas. CNEF provides a resource for harmonizing data ex-post, facilitating cross-national policy evaluation, increasing ex-ante harmonization in the panel setting, and expanding use of national data sets through repackaging.

¹⁶ Gruber, Jonathan and David Wise, eds. 1998. *Social Security and Retirement Around the World*. Chicago: University of Chicago Press.

Samuel Lucas

University of California, Berkeley
Associate Professor of Sociology

On Data Accessibility, Usability, and User Community Needs

Data Users

These comments are based on anecdotal experience from working with students and address the issue of the broader user community and use of the General Social Survey (GSS). Who are the users of GSS data? Potential data users (Table 1) may be roughly characterized as lay, novice and experts with respect to their knowledge and experience with data. Lay users are undergraduates and journalists searching for statistics, not using the data to generate knowledge to increase their understanding of a particular topic. The novice is students, journalists, and congressional staffers. Experts are those who repeatedly use the data set and are more familiar with data collection complexities and know statistical principles. This group increases the need for additional documentation and discussion about how to use the data, how it was gathered and what it means. Experts seek to disseminate analytic results and publish. Thus, as the proportion of experts grows, there is a greater proportion of those who are unable to access the data in a straightforward manner because of the hidden costs that come with ease of access to the data. The experts disseminate results that are likely subject to critical reassessment.

Given the knowledge level of data users, one may rightly be skeptical that much will be gained by making the raw data available to groups other than the experts. Data usage results from the level of difficulty involved in doing statistical analysis. Data users have more statistical firepower, but less ability to aim. It is just not possible for anyone to know everything about all data sets.

Table 1. A Rough Typology of Potential Data Users

	Lay	Novice	Expert
1. Repeated use of specific data set over multiple years	No	Maybe	Yes
2. Knowledge of data collection complexities	No	Maybe	Maybe
3. Knowledge of statistical procedures implementing principles	No	Maybe	Yes
4. Knowledge of statistical principles	Maybe	Maybe	Maybe
5. Regularly seeks to disseminate analytic results (e.g. publish)	No	Maybe	Yes
6. Disseminated results likely subject to critical reassessment	No	Maybe	Yes

Data Outreach Activities

Reaching expert and novice users will require different outreach activities (Table 2). The expert user would probably not attend workshops on older data sets or sessions by "outreach ambassadors" at more intimate conferences. Outreach at these types of events might be good locations to make contact with the novice user. Outreach at large professional association meetings are idea for outreach because graduate students, early career faculty and late career faculty all attend these conferences.

Outreach could go beyond the poster session at professional meetings by identifying persons who has worked with the data set and having them organize a regular session to discuss the data. Data outreach can also occur through media public relations (PR) with "splashy" research findings. This type of activity is important and might be of mutual interest for research authors and the GSS principal investigators (PIs). The PIs could contact the research author and see if they would be amenable to getting a small amount of support to raise the visibility of the data set. These news splashes could potential garner greater

public knowledge and support of the GSS ANES and PSID, which are very important for these databases and essential for their maintenance.

Table 2. Data set-specific Outreach Activities

Activities	Dilettante	Novice	Expert
1. Sessions or exhibits at major disciplinary conferences	No	Yes	Yes
2. Outreach "ambassadors" at more intimate conferences	No	Yes	No
3. Pre-or post-conference workshops on new data sets or releases	No	Yes	Yes
4. Pre-or post-conference workshops on older data sets	Maybe	Yes	No
5. E-newsletter announcing developments (e.g., planned releases)	No	Yes	Yes
6. Media PR with "splashy" research findings	Yes	Maybe	No

Data Depository Characteristics and User Needs

Table 3 provides a preliminary assessment of users needs by type of users (e.g., lay, novice, and expert) the type of access (i.e., data set-specific site and general data depository). Either access vehicle will meet the needs of the novice and expert users, but a general depository has the potential to reach and engage more dilettante users. Also, as a user or as a person teaching students, it would be more useful to go to onsite through ICPSR (Inter-university Consortium for Political and Social Research). There, however, are issues that should be considered in moving to a general depository. For instance, it is possible to turn the depositories to user portals, but it could be a risky endeavor. This would involve untimely changes of technology as well as upgrading technology, and massive resources would be required to maintain it. A general depository mode could be fine for unrestricted data, but presents problems with restricted data. This brings risks of spotty coverage across data sets. How would the funding structure be maintained? Large depositories must be maintained. The data specific sites are not unnecessary; to the user the data set sites justified.

Data users would need hardcopy searchable codebooks, general content on sampling, a graphic view of the instrument, the ability to extract specific content, and a graphic view of the extract in instrument. The data access system would need online preliminary statistic and graphic tools with cautionary pop-ups and links to teaching tools (see examples below). Expert users would also need access to restricted data. The graphic view of the instruments could tend to confusing unless developed with care. An example illustration is given below. The danger with the extract specific codebooks, is that it forces the user to access the data set and might not have the best measure for their question.

Figure 1. Examples of Cautionary Pop-ups

CAUTION: You are about to take the mean of a nominal variable. Statistical theory indicates this will likely be meaningless. If you want continue with this statistical operation, we advise you read LINK first, and, then check appropriate pop-up boxes after checking the box on the link.

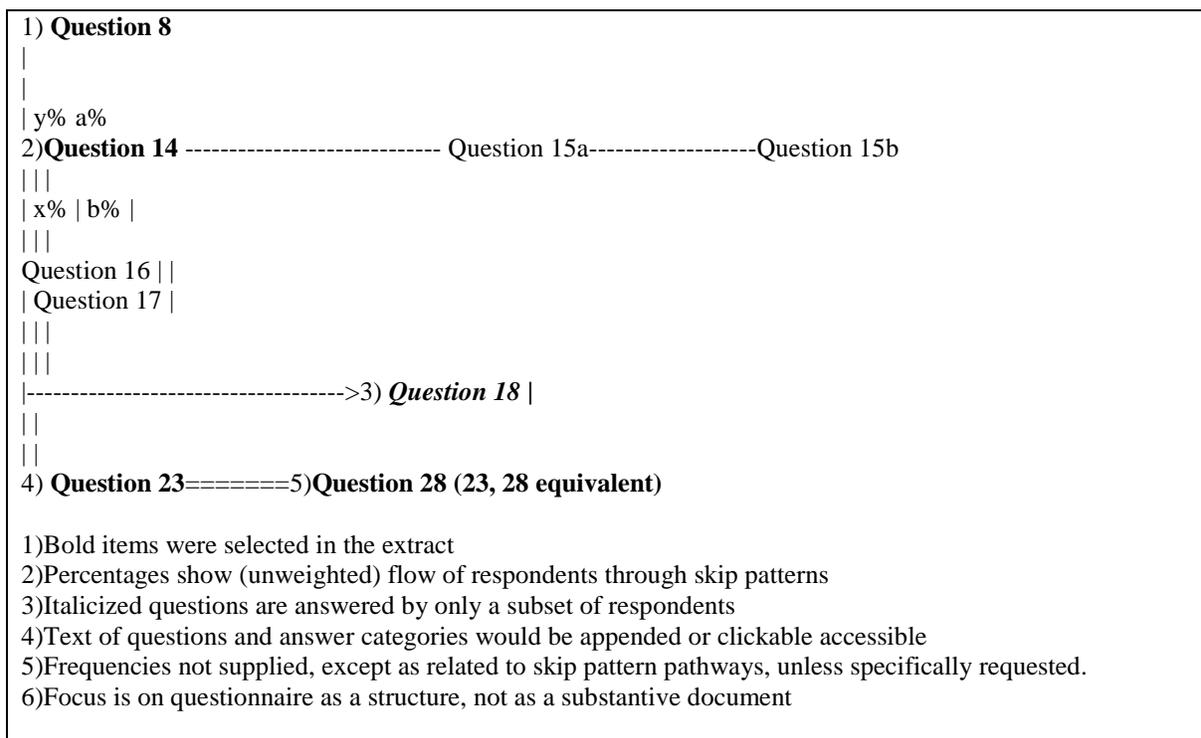
Or

DANGER: You are about to calculate a correlation coefficient between a nominal and an ordinal variable. If you really want to do this, check the next three pop-up boxes in the appropriate manner. Note that statistical theory suggests other measures of association would be more appropriate. To use those you will likely need to download the data, read it into a statistical analysis software package, and use the software to obtain the preferable measure of association. For more information on those measures, see LINK.

Table 3. A Preliminary Assessment of Depository Characteristics and User Needs

	General Depository			Data set-Specific Site	
	Lay	Novice	Expert	Novice	Expert
1. Unrestricted data	Yes	Yes	Yes	Yes	Yes
2. Restricted data	No	No	No	Novice	Yes
3. Hardcopy (searchable pdf) codebook	Yes	Yes	Yes	Yes	Yes
<i>a. General Content</i>	Yes	Yes	Yes	Yes	Yes
<i>b. Graphic View of Instrument</i>	Yes	Yes	Yes	Yes	Yes
<i>c. Extract Specific Content</i>	Yes	Yes	Yes	Yes	Yes
<i>d. Graphic View of Extract in Instrument</i>	Yes	Yes	Yes	Yes	Yes
4. Raw data	No	Yes	Yes	Yes	Yes
5. Online preliminary statistics tools	Maybe	Yes	Yes	Yes	Yes
<i>a. Cautionary pop-ups</i>		Yes	Maybe	Yes	Yes
<i>b. Links to methodological training materials</i>		Yes	Maybe	Yes	Yes
6. Online preliminary graphing tools	Maybe	Yes	Yes	Yes	Yes
<i>a. Cautionary pop-ups</i>		Yes	No	Yes	No
<i>b. Links to methodological training materials</i>		Yes	No	Yes	No

Figure 2. Example of Structure of Graphic View of Instrument



Lynn Smith-Lovin
Duke University
Robert L. Wilson Professor of Arts and Sciences

Future Investments in Survey Data Access & Dissemination

These comments revolve around the value-added of the General Social Survey (GSS), American National Election Study (ANES) and Panel Study of Income Dynamics (ANES) from the perspective of a teacher or someone training many users. It should first be noted that there are things the different projects are doing very well in an effort of best practices, but there are still developing trends and challenges for future investments.

Teaching: Dissemination for Dummies

New data users are not really "dummies," but are the scientists of tomorrow. Part of the job of the teachers and NSF is to excite future students and new data-driven knowledge is a good approach to learning. The GSS, ANES and PSID data sets are used to teach tomorrow's scientists in statistics and methodology courses. They are particular great teaching tools for students in a first year seminar because they allow students to generate their own research questions, analyze data and to answer their questions. So there is a real place in teaching for the data from the three projects.

Survey Documentation and Analysis¹⁷ (SDA) is generally used to teach students and it is a great analysis tool. Teaching would be enhanced by adding data sets and by creating a student version of SDA that is even simpler. The version would have imputed data, common controls, and not include narrowly useful variables such as vignette designs. Now it takes considerable implicit knowledge to learn how to analyze the data using SDA. It would be much easier if variable labels (to make use much easier), flat files (for all users), usual-suspect control variables, and variable universe variation were provided to make SDA even less complicated to use.

In teaching students data analysis, training students how to ask appropriate questions and how to approach answering them is a big task. Then teaching them concepts such as spurious relationships, control variables, and missing and imputed data takes a big effort. While making these available might make the data unusable for the expert community, a standard set of control variables could create a good teaching tool. For example, it is not useful to have the "race" variable in 231 categories, but having a canned set of control variables is a very valuable training tool, making the entry to data analysis much easier.

Dissemination for Principal Investigators (PIs)

The three data projects are valuable, but there are opportunities to enhance dissemination and usability. The GSS would benefit users by providing bibliographic links through variable use and being more readily accessible through more portals. The PSID could provide standard data sets (many will want to use), making it possible for students to ask questions to get them asking the right questions, Add Health could develop user conferences to improve the low number of users that can use the data optimally. It would be beneficial to bring users together to share their knowledge. The ANES could increase users by getting the data out faster and then disseminating the full data.

¹⁷ SDA is a set of programs for the documentation and Web-based analysis of survey data. SDA is developed and maintained by the Computer-assisted Survey Methods Program (CSM) at the [University of California, Berkeley](http://www.berkeley.edu). CSM also develops the [CASES](http://www.cases.berkeley.edu) software package

Future Issues

Future challenges for data access and dissemination lie in multi-level data sets, particularly hypernetwork relational data sets. Another issue is the adding of more contextual data, for example geo-codes. Issues surrounding the creation of standardized codes and sampling systems, ease of data linkage across data sets and IRB (institutional review boards) issues have to be thought through and addressed. Data providers and data access must be sensitive to IRB issues, for instance, being attentive to the assurances promised with using sensitive data and secure research environments. As more young scholars are using R and other open access software what are the implications for data sets, code for handling missing data, recodes, and archiving of published analyses? Decisions about standards for these issues should be communicated across the user community so that the wheel does not need to be reinvented multiple times. Finally, there is a need to work harder to find ways to archive the published analyses so that it could be replicated, challenged and understood.

Jennifer Stoloff

Office of Policy Development and Research
Housing and Urban Development (HUD)
Social Science Analyst

Potential for Non-Academic Use of PSID Data

I am representing a different community; economic PhDs go into other employment sector, for example the government, not just academics. Of the three data projects-- Panel Study of Income Dynamics (PSID), General Social Survey (GSS) and American National Election Studies (ANES) --I am most familiar with the PSID. I used the PSID in my dissertation research and at one time the Department of Housing and Urban Development (HUD) participated in the funding of the PSID.

When I first started to work with the PSID data, it was not as accessible as it is now. I had to download each year of data and print out the large 700 page codebook. Access to the longitudinal data was even more difficult. Access to the PSID data has significantly evolved since then and access to the longitudinal data is now also much easier.

At the present time, HUD supports the PSID more in theory than actual in practice. At HUD most of the analysts who could potentially complete useful policy analysis from PSID do not have time to conduct independent analyses with data sets like the PSID. They rely upon much of HUD's own internal administrative data, and are thus constrained in the questions that can be asked. For example, most HUD data will always not include measures of education. In collecting data HUD may ask questions about public income and welfare assistance, and information to calculate the rent that residents can be charged, but cannot ask about education. If it was possible to link the HUD data to PSID data on education this would be extremely useful. The PSID contains questions about subsidized housing, but these data are not very reliable. There, however, are significant confidentiality issues around linking HUD and PSID data. There have been some attempts to match these data, but they were not very successful

The focus of HUD analysts is somewhat limited and narrow in that their analyses must contribute to the Department's mission, and using PSID would involve considerable staff time to make the PSID accessible and useable. The PSID does a good job of outreach, but HUD staff cannot make good use of the data because of time constraints. There is a plan to continue the work to keep ongoing access to PSID data, but there are IRB (institutional review board) issues that have to be resolved for a federal user to get access to the geo-coded data. There is an IRB now at HUD, but it has not been used very much and not many staff know about it requirements and existence. This is probably good since HUD is not a statistical agency and HUD (and thus access to the PSID) data is not shared with independent researchers.

Looking forward, the challenge for HUD is to identify a question or issue on which it needs to be informed and the answer it is in the PSID. It would be great if the relevance of the PSID and GSS data for policy research could be determined and expanded to increase major usability at the governmental level. The PSID and GSS data could potentially provide useful measures for HUD so that they do not have to complete their own data collection.

Mark Wilhelm

Indiana University-Purdue University Indianapolis
Board of Directors for the Panel Study of Income Dynamics (PSID)

Using ANES, GSS, and PSID Data

I am a user of the Panel Study of Income Dynamics (PSID), General Social Survey (GSS) and American National Election Studies (ANES). There are two metaphors--telescopes and microscopes--that can be applied to the PSID, GSS and ANES. Constructing data sets and using data is like making bread. Data preparation is similar to making bread in that 80 percent of the time is spent making it and 20 percent eating it; similarly, 80 percent of the time is spent preparing data and 20 percent analyzing it. It is striking, however, how little time is spent in "exchanging recipes" especially when often many are baking the same bread. Given that the planned outcome of data preparation is data analysis it is amazing that there is not more training for graduate students in data analysis.

The PSID is a good illustration since of the three data sets it is the most complex to use. A few useful models for building user collaboration, for example, Dean Lillard's collaboration processes (see page 69-72 of this report) have been presented. The PSID has variables over time and they have to be harmonized overtime. The ANES and GSS actually measure variables, but with the PSID, there are elements that have to be combined to create a variable like, for example, income. In the case of income, the PSID creates a generated variable that is available, but there are plenty of cases where this is not the situation, that is variables are not generated. The user has to spend time to do so. If using STATA and accessing the journal, there are some sub routines within STATA whereby the user can download the code as it is or find something that is close to what they want to use and amend it to do what they want. It would be extremely useful if the PSID could do this using a format like a PSID journal for users. The journal would have articles describing how a variable, for instance, total family income has been created, how it has been harmonized across the years and provide a description of the metadata. The metadata would make the version of the processes visible, including identification of the version of the release of the PSID. The user could download a small data file with the variables and a quick help file describing the variable and the code. The name of the variable might solve the PSID naming problem. The income variable has been constructed, but the health variables, charitable giving and others have not. The journal could provide a format to discuss the ways in which the complex PSID survey design can be used, along with the weights.

Having this kind of PSID journal would provide benefits for first, the expert users. The major benefit would be time saved in completing an analysis project. The journal would explain how the variable was built by the user, so that other users could later use it. This would save the users time, especially for variables that take enormous time to construct such as consumption expenditures. A journal would lead to efficiency and potentially rapid scientific discovery.

If the PSID is made easier to use by alleviating the need for users to rebuild variables, it is then easier for people to move between the data sets. They would not have to recreate the variables for analysis and could shift their research agendas from data driven to question driven. This also would encourage cross-discipline work, lower entry costs for new users, increase the replicability of previous work and allow users to judgments they often make in their data analysis. The journal could contain columns for less sophisticated users with one or two variable data sets that could be used as teaching tools. The PSID strength is that it is transforming the views of the world from ones that are static to more dynamic views of the world, but this transformation has not reached the undergraduate level. The PSID could publish data briefs similar to the briefs at NSF, focusing on the discoveries in the data. The main idea is to create a forum so there is more time for innovation and less time spent with preparation for completing analysis.

In the case of the GSS and ANES, the GSS is easy to use but the coding for religious affiliation and denominations is difficult to do. A GSS journal with the documentation of how the codes could be combined would be helpful to users. The ANES has an implicit association test and many people would not know how to get it themselves. Many different approaches are used for graduate training and data preparation and there are best practices. These need to be more broadly disseminated. Journals would formalize the use of the data and also assist with the ongoing citation and metadata issues.

APPENDIX 3 – Presenters' Biographical Sketches

George Alter

George Alter is Acting Director of the Inter-university Consortium for Political and Social Research (ICPSR), Research Professor at the Population Studies Center, and Professor of History at the University of Michigan. ICPSR is the world's largest social science data archive, and it includes units that specialize in data on aging, childcare, criminal justice, demography, health, and substance abuse. Alter is also principal investigator for an NSF award creating a pathway within the National Science Digital Library to promote the use of social science data in teaching quantitative literacy. His research grows out of interests in the history of the family, demography, and economic history, and recent projects have examined the effects of early life conditions on health in old age and new ways of describing fertility transitions. Recent publications include: Alter and Oris, "Effects of Inheritance and Environment on the Heights of Brothers in Nineteenth-century Belgium." *Human Nature* (2008); and Alter, Dribe, and Poppel, "Widowhood, Family Size, and Post-Reproductive Mortality: A Comparative Analysis of Three Populations in Nineteenth Century Europe," *Demography* (2007).

Peter Bearman

Peter Bearman is the Director of the Lazarsfeld Center for the Social Sciences, the Cole Professor of Social Science, and Co-Director of the Health & Society Scholars Program at Columbia University. He was the founding director of the Institute for Social and Economic Research and Policy at Columbia, serving from the Institute's launch in 2000 until 2008. A recipient of the National Institutes of Health Director's Pioneer Award in 2007, Bearman is currently investigating the social determinants of the autism epidemic.

A specialist in network analysis, he co-designed the National Longitudinal Study of Adolescent Health and has used the data extensively for research on topics including adolescent sexual networks, networks of disease transmission, and genetic influences on same-sex preference. He has also conducted research in historical sociology, including *Relations into Rhetorics: Local Elite Social Structure in Norfolk, England, 1540-1640* (Rutgers, 1993). He is the author of *Doormen* (University of Chicago Press, 2005).

Andrew Beveridge

Andrew A. Beveridge, Ph.D., is Professor of Sociology at Queens College and the Graduate School and University Center of CUNY. He chairs the Queens College Sociology Department. Since 1993, Dr. Beveridge has been a consultant to the NY Times, which has published numerous news reports and maps based upon his analysis of the Census data. He writes the demographic topic column for the Gotham Gazette, an on-line publication of the Citizens Union. He is working on three major projects involving urban and neighborhood change, one tracking long term change in US urban areas and two that look at neighborhood impact on individuals, including drug users and drug dealers, high school and elementary school students, and economic and other standing. He is also collaborating on a study of test score patterns in the Houston Independent School District. He collaborates with a team of researchers at the University of Minnesota and other institutions on the National Historical Geographic Information System project, which is producing data to examine long term trends in major cities and urban areas in the US.

Dr. Beveridge is an expert in using GIS (geographic information system) techniques to integrate demographic materials. He and his team have developed an interactive application and Web Based set of maps entitled Social Explorer (www.socialexplorer.com) that allow the user to compare and contrast demography based upon an area that he or she selects. This work was funded by NSF and the New York Times. He has examined the social roots of American banking and credit practices; public attitudes towards science and technology; factors leading to union success in winning representation elections; social trends revealed by housing surveys; and economic development in Africa. He is the co-author of *African Businessmen and Development in Zambia*, Princeton University Press, and numerous articles, papers and reports. His research work has received grant and fellowship support from the American Council of Learned Societies, the NSF, the National Endowment for the Humanities, the American

Philosophical Society, the Department of Housing and Urban Development, the Putnam Foundation, the Robert Wood Johnson Foundation and other agencies.

Mark Chaves

Mark Chaves is Professor of Sociology, Religion, and Divinity at Duke University. Among other projects, he directs the National Congregations Study (NCS), a wide-ranging survey, conducted in 1998 and again in 2006-07, of a nationally representative sample of religious congregations. NCS results have helped us to better understand many aspects of congregational life in the United States. Professor Chaves is the author of *Congregations in America* (Harvard, 2004), *Ordaining Women: Culture and Conflict in Religious Organizations* (Harvard, 1997) and many articles, mainly on the social organization of religion in the United States. His in-progress book on religious trends in the United States is under contract with Princeton University Press. He currently is Chair of the General Social Survey's Board of Overseers and immediate Past-President of the Society for the Scientific Study of Religion.

Matthew DeBell

Matthew DeBell is Director of Stanford Operations for the American National Election Studies. He manages the survey development and data collection and dissemination activities for ANES at Stanford and served as study director for the 2008-09 Panel Study and the 2006 Pilot Study. He contributes to questionnaire development, manages contracts for data collection, oversees field operations, reviews and designs survey methods, develops quality assurance procedures, and documents and analyzes data. He also supervises Stanford staff, research assistants, and subcontractors. Before joining the ANES staff in 2006, Matt worked as a senior research analyst for the American Institutes for Research in Washington, DC, where he contributed to the National Household Education Surveys Program and the Early Childhood Longitudinal Studies for the Department of Education. He received his Ph.D. in government from Georgetown University.

Darrell Donakowski

Darrell Donakowski is the Director of Studies for the American National Election Studies (ANES) He is responsible for project management, oversight of data collection and outreach activities, protection of archived study materials, and setting direction, timelines, and staffing levels for operations. Darrell works with the PIs and the ANES Board to set project goals, maintain linkages between studies in the time series, and ensure continuity in the ANES program of research. Darrell also has oversight of financial administration, post-award activity, personnel matters, and prepares reports to funding organizations. He works with the PIs and the ANES Board to make decisions about study design, field operations, and dissemination. Prior to joining ANES in 2008, he was a project manager at ICPSR for the Data Preservation Alliance for the Social Sciences (Data-PASS), a broad-based partnership devoted to identifying, acquiring and preserving data at-risk of being lost to the social science research community. Darrell received his M.A. in Social Psychology from the University of Western Ontario.

Pamela Herd

Pamela Herd is Associate Professor of Public Affairs and Sociology at the University of Wisconsin-Madison. After receiving a PhD in Sociology from Syracuse University Pamela became a Robert Wood Johnson Foundation Scholar in Health and Health Policy at the University of Michigan, Ann Arbor. Her research examines the relationship between education, income, and health. Herd is author of numerous articles and chapters that have appeared in *Social Forces*, *Gender and Society*, *Journal of Health and Sociology Behavior*, *Psychosomatic Medicine*, *Milbank Quarterly*, *Journals of Gerontology* and *The Gerontologist*. Pamela is currently the Co-PI, with Robert Hauser, of the Wisconsin Longitudinal Study.

David Howell

David Howell is the Assistant Director of the Center for Political Studies at the University of Michigan's Institute for Social Research. He has been the Director of Studies and organized the Secretariat for the Comparative Study of Electoral Systems (CSES; www.cses.org) since 2001. He worked previously for seven years on the American National Election Studies (ANES; www.electionstudies.org) and eight years on the Health and Retirement Study (HRS; www.hrsonline.isr.umich.edu), including participating in the development and implementation of dissemination and archiving strategies for both. In addition to research and administration, he has skills and experience in the areas of programming, technology, and website development.

Michael Hout

Michael Hout holds the Natalie Cohen Sociology Chair at the University of California-Berkeley. He teaches courses on inequality, data analysis, and population. In his research, Mike uses demographic methods to study social change in inequality, religion, and politics. In 2006, Mike and Claude Fischer published *Century of Difference*, a book on twentieth-century social and cultural trends in the United States that exemplifies this approach. Another book, *The Truth about Conservative Christians* with Andrew Greeley (University of Chicago Press, 2006) is another example. A couple of illustrative papers include "How Class Works: Subjective Aspects of Class Since the 1970s" in a book edited by Annette Lareau and Dalton Conley (Russell Sage Foundation 2008), "The Demographic Imperative in Religious Change" (*Am. J. of Soc.*, Sept. 2001) and "How 4 Million Irish Immigrants Came to be 40 Million Irish Americans" (with Josh Goldstein, *American Sociological Review*, April 1994). Previous books are: *Following in Father's Footsteps: Social Mobility in Ireland* (Harvard Univ. Press 1989) and, with five Berkeley colleagues, *Inequality by Design* (Princeton Univ. Press, 1996). Mike Hout's honors include election to the American Academy of Arts & Sciences in 1997, the National Academy of Sciences in 2003, and the American Philosophical Society in 2006. Mike currently chairs the Graduate Group in Sociology and Demography and the Berkeley Population Center. Mike's education includes a bachelor's degree from the University of Pittsburgh in history and sociology and a master's and doctorate from Indiana University in sociology. He taught at the University of Arizona before coming to Berkeley in 1985.

Vincent Hutchings

Vincent Hutchings is a Professor of Political Science at the University of Michigan and a Research Professor at the Institute for Social Research. He received his Ph.D. in 1997 from the University of California, Los Angeles. Professor Hutchings teaches courses in African American politics, public opinion and voting behavior, and legislative behavior in the US. His research interests focus on the circumstances under which citizens are attentive to political matters and engage in issue voting. He published a book on this topic entitled *Public Opinion and Democratic Accountability* in 2003, from Princeton University Press. His research also examines the ways in which political appeals carried through the mass media can influence attitudes about salient social groups such as racial and ethnic minorities and women. Additionally, his work has explored the ways in which political campaigns can frame information about racial issues in order to activate and make politically relevant voter attitudes about particular racial groups. His current project focuses on inter-racial and inter-ethnic competition and the ways in which elite communications can exacerbate or diminish inter-group conflict. His work has appeared in the *American Sociological Review*, the *American Political Science Review*, the *Journal of Politics*, the *Annual Review of Political Science*, *Political Communication*, *Public Opinion Quarterly*, *Political Psychology*, the *Journal of Communication* and *Legislative Studies Quarterly*. Professor Hutchings has received multiple grants from the National Science Foundation, most recently (2009) for his project entitled "Elite Communications and Racial Group Conflict in the 21st Century." In 2004, he served as co-Principal Investigator of the National Politics Study, a national survey of Whites, Latinos, African Americans, Afro-Caribbeans and Asian Americans. From 2000-2002 he was a Robert Wood Johnson Foundation Health Policy Scholar at Yale University. He served on the American National Election Study (ANES) Board from 2005-2009, and also took on the role as Associate Principal Investigator of the study from 2007-2009.

Simon Jackman

Simon Jackman's research centers on American electoral politics, public opinion, democratic representation and the art and science of survey research. In recent years his research has investigated the Internet as a platform for survey research, to better track the evolution of public opinion and produce more politically relevant assessments of American political attitudes. In 2007-08 he was one of the principal investigators of the Cooperative Campaign Analysis Project, an Internet-based, six-wave, longitudinal study of the American electorate leading up to the 2008 Presidential election. Jackman co-directs the Stanford Center for American Democracy and is one of the principal investigators of the American National Election Studies, 2010-2013. Jackman also co-directs the Methods of Analysis Program in the Social Sciences at Stanford. In 2009 he published a 600 page statistics text, *Bayesian Analysis for the Social Sciences* (Wiley). Jackman is also an associate editor of *Annual Reviews of Political Science* and *Political Analysis*. Recent articles have appeared in the *American Journal of Political Science*, the *Journal of Politics*, and the *Journal of Elections, Public Opinion and Parties*. Jackman has taught at Stanford since 1997; he taught at the University of Chicago 1994-96 and holds a PhD from the University of Rochester (1995).

Dean Lillard

Dean Lillard received his PhD in economics from the University of Chicago in 1991. He has been a member of the Department Policy Analysis and Management at Cornell University since 1991. He is currently Senior Research Associate and Co-Director and Project Manager of the Cross-National Equivalent File (CNEF) study. The CNEF project compiles and equalizes data from household panel studies in seven countries. The data are distributed to researchers worldwide. He is also a Research Associate at the German Institute for Economic Research in Berlin.

Dean Lillard's current research focuses on health economics, the economics of schooling, and international comparisons of economic behavior. His research in health economics is primarily focused on the economics of cigarette marketing and consumption. His research on the economics of schooling includes studies of direct effects of policy on educational outcomes and on the role that education plays in other economic behaviors such as smoking, production of health, and earnings. His cross-national research ranges widely from comparisons of the role that obesity plays in determining labor market outcomes to comparisons of smoking behavior cross-nationally.

Dean Lillard is a member of the American Economics Association, the American Society of Health Economists, the Population Association of America, the Association for Public Policy Analysis and Management, the International Association for Research on Income and Wealth, and the International Health Economics Association. He also sits on the Advisory Board of the Danish National Institute for Social Research.

Nimmi Kannankutty

Nimmi Kannankutty is a senior advisor in the National Center for Science and Engineering Statistics at the National Science Foundation. She has responsibility for outreach and dissemination activities, and also spends considerable time on analytic work using NCSES and other large-scale data sets, including being the lead author of the workforce chapter of the Science and Engineering Indicators report. Prior to her current position, she had been the coordinator for NSF's three workforce surveys of scientists and engineers, which make up the Scientists and Engineers Statistical Data System, or "SESTAT". Dr. Kannankutty's areas of science policy expertise include the science and engineering workforce, graduate education and academic research; she has also been trained in survey and statistical methods.

Samuel Lucas

Samuel R. Lucas, an Associate Professor of Sociology at the University of California-Berkeley, has research and teaching interests in social stratification, sociology of education, research methods, and research statistics. His most recent article, published in *Rationality and Society* in October 2009, translated multiple theories of inequality into equations, and, using the equations, critically assessed the theories' equivalence and falsifiability. Notably, he found that two ostensibly conflictual theories have interesting points of agreement, and another widely replicated theory of inequality in education research, maximally maintained inequality, cannot be falsified. Publishing multiple articles on educational attainment, he recently edited a special issue of *Research in Social Stratification and Mobility* on new developments in education transitions research. His book, *Tracking Inequality: Stratification and Mobility in American High Schools*, received the Willard Waller Award from the Sociology of Education Section of the American Sociological Association in 2000 for the most outstanding book in the sociology of education for 1997, 1998, and 1999. His most recent book, published in 2008, is *Theorizing Discrimination in an Era of Contested Prejudice*, the first of a 3-volume series on race and sex discrimination in the United States.

Peter Marsden

Peter V. Marsden is Harvard College Professor and Edith and Benjamin Geisinger Professor of Sociology at Harvard University. He received his undergraduate degree (Sociology and History) at Dartmouth College (1973) and his graduate degrees (Sociology, MA [1975] and Ph.D. [1979]) at the University of Chicago.

He has been a Co-Principal Investigator of the General Social Survey (GSS) since 1997. From 1988 until 1997 he served on the GSS Board of Overseers, chairing that board from 1993 until 1997. He directs the Program on Survey Research at Harvard within the Institute for Quantitative Social Science. Marsden's substantive research interests center on social organization, especially formal organizations and social networks; they also involve the sociology of medicine. He has ongoing interests in social science methodology and survey research. He edited the 1991-1995 volumes of *Sociological Methodology*, and is the co-editor (with James D. Wright) of the second edition of the *Handbook of Survey Research* published by Emerald Group Publishing in May, 2010. He has examined aspects of the measurement of social networks using survey methods for many years; a forthcoming chapter on this subject will appear in the *Sage Handbook of Network Analysis*. Apart from his work on the GSS, Marsden was a lead investigator of three National Organizations Studies conducted between 1991 and 2003, and used those establishment survey data in investigations of organizational factors linked to the presence of various human resource practices—recruitment and staffing practices, “high performance” practices, and the use of “contingent” workers—in U.S. workplaces. He is currently in the last stages of assembling a collection of studies of U.S. social trends since 1972 based on GSS data. Marsden teaches courses on quantitative methods, research methods, organizational analysis, and social networks. He served on the advisory review panel for the Sociology Program in the National Science Foundation from 1983 until 1985.

Timothy Mulcahy

Timothy Mulcahy is a Senior Research Scientist at the National Opinion Research Center (NORC) at the University of Chicago and Program Director of the NORC Data Enclave. He has nearly 20 years experience in social science research and is a leader in promoting international adoption of data and metadata standards, best practices in data documentation, and digital age dissemination. He has served as an invited speaker and panelist at numerous conferences and workshops focusing on microdata access, confidentiality, disclosure avoidance, and dissemination. He also serves as Co-Principal Investigator of a grant jointly sponsored by the National Institute of Justice (NIJ) and the National Institute on Drug Abuse (NIDA) examining illicit retail drug markets across the U.S. Mulcahy earned his graduate degree from the Institute for Policy Studies (IPS) at the Johns Hopkins University; conducted his international economics and foreign policy studies at the Johns Hopkins University Paul Nitze School of Advanced International Studies (SAIS); and earned his undergraduate degree from the University of Virginia.

Steven Ruggles

Steven Ruggles is Regents Professor of History and Population Studies, Distinguished McKnight University Professor, and Director of the Minnesota Population Center at the University of Minnesota. He designed and developed four large-scale data infrastructure projects: the Integrated Public Use Microdata Series (IPUMS), the Integrated Health Interview Series, and North Atlantic Population Project. He has been awarded 47 grants with over \$80 million total costs by NSF and NIH to support these projects, and has received the William J. Goode Award from the Family Section of the American Sociological Association, the Allen Sharlin Award from the Social Science History Association, the Robert J. Lapham Award from the Population Association of America, and the Warren Miller Award from the Inter-university Consortium for Political and Social Research. He has published extensively in the field of historical family demography, especially long-run changes in intergenerational coresidence, living arrangements of children, divorce and separation, fertility, migration, demographic methods, and the development of population data infrastructure.

Narayan Sastry

Narayan Sastry is a Research Professor in the Population Studies Center and Survey Research Center and Associate Director of the Survey Research Center at the University of Michigan's Institute for Social Research. He is also an Adjunct Senior Social Scientist at the RAND Corporation. Sastry has been at the University of Michigan since 2006. Sastry was previously a Senior Social Scientist at RAND and Associate Director of RAND's Labor and Population Program and Population Research Center. Sastry's research interests center on studying the social and spatial dimensions of health, development, and well-being of children and adolescents, both in the United States and in less developed countries. Sastry is the Co-Director of the Los Angeles Family and Neighborhood Survey (L.A.FANS). He is the Director of the Displaced New Orleans Residents Survey (DNORS) that is being designed to study the long-term demographic effects of Hurricane Katrina on the pre-storm population of New Orleans. Sastry also serves as a Co-PI on the Panel Study of Income Dynamics (PSID). Sastry received his Ph.D. in Demography and Public Affairs from Princeton University in 1995.

Lynn Smith-Lovin

Lynn Smith-Lovin is Robert L. Wilson Professor of Arts and Sciences in the Department of Sociology (with secondary appointments in Psychology and Neuroscience and in Women's Studies) at Duke University. She received the 2006 Cooley-Mead Award for lifetime achievement in social psychology from the American Sociological Association Section on Social Psychology and the 2005 Lifetime Achievement Award from the Section on Sociology of Emotions. Her research examines the relationships among identity, action and emotion. Her current projects involve (1) an experimental study of justice, identity and emotion (funded by the National Science Foundation), (2) research with Miller McPherson on an ecological theory of identity (also funded by NSF), and (3) a study of event processing in Arabic (funded by the Office of Naval Research). She has served as President of the Southern Sociological Society, Vice-President of the American Sociological Association, and Chair of the ASA Sections on Social Psychology and the Sociology of Emotion.

Tom Smith

Tom W. Smith is an internationally recognized expert in survey research specializing in the study of societal change and survey methodology. He is Director of the Center for the Study of Politics and Society at the National Opinion Research Center, University of Chicago. Since 1980 he has been a principal investigator of the National Data Program for the Social Sciences and director of its General Social Survey (GSS). He is also co-founder and former Secretary General (1997-2003) of the International Social Survey Program (ISSP). The ISSP is the largest cross-national collaboration in the social sciences.

Smith has taught at Purdue University, Northwestern University, the University of Chicago, and Tel Aviv University. He was awarded the 1994 Worcester Prize by the World Association for Public Opinion

Research (WAPOR) for the best article on public opinion, the 2000 and 2003 Innovators Awards of the American Association for Public Opinion Research (AAPOR), the 2002 AAPOR Award for Exceptionally Distinguished Achievement and the Eastern Sociological Society Award for Distinguished Contributions to Sociology in 2003. He is currently President-Elect of the World Association for Public Opinion Research.

Jennifer Stoloff

Jennifer Stoloff has worked as a Social Science Analyst for HUD since 1999 in the Program Evaluation Division in the Office of Policy Development and Research. Her research responsibilities include studies on HUD's multifamily housing, the Family Self Sufficiency Program, and soon further research about Native American Housing Needs and Rent Reform. Jennifer earned her PhD in Sociology at the University of North Carolina at Chapel Hill, specializing in demography and urban sociology. Her dissertation project, which was about public housing and labor force participation, included a longitudinal analysis of the PSID. Because of that experience, Jennifer has served as HUD's representative to the PSID Board of Overseers since coming to HUD.

Lois Timms-Ferrara

Lois Timms-Ferrara is Associate Director at the Roper Center for Public Opinion Research at the University of Connecticut. She has been with the Center for more than 25 years, serving in various research and administrative capacities. She has coordinated large scale research projects and managed user and data services. She is responsible for Center outreach programs, data acquisitions, membership development, and plays a key role in identifying functionality to enhance access features for data delivery.

Over the last five years, Lois has been active in the data rescue efforts of the Data-PASS project. Data-PASS, partially funded by the Library of Congress, is a partnership of major social science archives with a mission to identify, locate, preserve and make accessible critical social science data that has yet to be archived. Her efforts have focused on locating and archiving the USIA collection of data spanning 1952 to 1999, and studies that remain in punch card format conducted by the National Opinion Research Center in the 1950s and 1960s. Her current interests lie in developing educational tools for instruction in the use of polling data and the development of high school curriculum tools that introduce the concept of data analysis into secondary schools. Her educational background is in economics and organizational relations.

Mark Wilhelm

Mark O. Wilhelm is Professor of Economics and Philanthropic Studies at Indiana University-Purdue University Indianapolis and conducts empirical research on prosocial behavior. He is the Founding Director of the *Center on Philanthropy Panel Study*, a project that since 2001 has gathered data on giving and volunteering through a module on the *Panel Study of Income Dynamics*. His most recent analyses of these data are about the intergenerational transmission of generosity, experiences of family instability/low income during adolescence and subsequent giving/volunteering in young adulthood, and the relationship between religious affiliation and giving to organizations that help people with basic needs. Professor Wilhelm has prepared user-friendly extracts of the *PSID* giving and volunteering data intended to facilitate wider use.

Professor Wilhelm has used the *General Social Survey* for several projects, most recently for an article examining the associations between empathic concern and a moral principle of care with prosocial behavior. He designed the Giving and Religion module on Wave 22 of the 2008-2009 *ANES Panel Study*. Professor Wilhelm's other recent research is experimental work on altruistic motivation and econometric work on the use of specification tests in censored regression models. His earlier prosocial behavior research dealt with voter support for public assistance to the poor and help given within the family.

