Data Access and Archiving: Options for the Demographic and Behavioral Sciences Branch

Lori Melichar, Jeffery Evans, and Christine Bachrach

August 2002

Summary:

The primary objective of the DBSB of the National Institute of Child Health and Human Development (NICHD) is to advance the field of population research. In the past, the Branch has contributed to this objective by funding innovative population research and data collection efforts. DBSB- and NICHD-funded datasets, Datasetsuch as the National Longitudinal Study of Adolescent Health (Add Health), the National Survey of Families and Households (NSFH), and the Integrated Public Use Micro Surveys (IPUMS) datasetare useful for answering a wide range of empirical questions that had previously seemed out of reach. Challenges associated with sharing sensitive data have prompted the principal investigators (PIs) of these data collection projects to develop data-sharing processes that are now commonly accepted as exemplars in the field. As the amount of information that can be collected and analyzed about an individual further multiplies, future technological, cultural, and legal developments will force future PIs to seek out even more effective ways to protect and share data. These conventions will not only

enable the population research community to produce a large number of secondary analyses of these data, but will also allow the larger research community to take advantage of unique data resources.

To continue supporting the field of population science, the DBSB plans to sustain current successful data collection efforts, support innovative new surveys, and seek and enable implementation of new ways to increase access to data collected with National Institutes of Health (NIH) funding. The DBSB convened a panel of "expert" data collectors, archivers, marketers, and users in fall 2001 to discuss experiences and challenges associated with these tasks. Jeff Evans, the project officer who convened the workshop, posed the following questions to the panelists for discussion:

1. How should the DBSB help preserve long-term access to scientific data when funding for the original project has expired?
2. How can the DBSB best assist investigating teams in meeting the increasingly complex legal requirements involved in data sharing?
3. How can the DBSB ensure that researchers continue to have access to data of increasing sensitivity?
4. How can the DBSB ensure that large-scale data collection projects have sufficient funding to sustain adequate access?

The DBSB asked workshop attendees to suggest ways to improve the general level of sophistication that exists in the population research community regarding data sharing.

Panelists urged the DBSB to invest thought and resources into finding ways to make the data archiving and access system more efficient, effective, and sustainable.

In the pages that follow, we first discuss current data sharing standards, options for data distribution and storage, and confidentiality issues associated with the data-sharing process. We then discuss past, current, and future challenges, many of which have had our attention for some time, while others were brought to our attention by the conference participants. The conclusion of this paper presents a list of proposals to address panelists' expressed need for resources and guidance to enable PIs and archivists to work together so that social scientists and others may have access to confidential data.

Data Sharing Standards: Access, Preservation, and Confidentiality

The DBSB feels strongly that researchers should provide access to data in order to increase their impact on the research community. The "minimum data-sharing expectation" held by most of the population research community consists of three principles: 1) a reasonably complete data file should be released within a reasonable amount of time[1] after the PI has completed data collection; 2) the data should be stored after grant support expires; and 3) the confidentiality of information provided by human subjects must be preserved[2]. Similar principles have recently been incorporated into

[1] For most kinds of data collection activities, the expectation is that a reasonably complete data file should be released within a year of the completion of data collection. Data that require extraordinary effort at the coding stage are expected to be released within two to three years.

[2] The process by which DBSB-supported data are shared with the scientific community is exemplified by the NSFH, a study that has been widely praised for its data-sharing practices. NSFH researchers released a clean version of the study dataset via the Web or CD-ROM for free or a nominal charge no later than one year after the data are gathered from the field. The PIs supervised the process of data cleaning and variable construction, processes that often continue after the initial data release. In addition to releasing data directly to users, the NSFH PIs also transferred the dataset to an established archive for preservation and distribution in user-friendly formats.

proposed guidelines for data sharing for all NIH-supported research[3]. These guidelines encourage all researchers with NIH support who are collecting data[4] to consider making the data available for broader use and state that shared data should be "free of identifiers that would permit linkages to individual research participants, and variables that could lead to deductive disclosure of individual subjects."[5]

Providing widespread, timely access to data increases the benefit to society on many levels. Expedition of data cleaning, provision of technical support, and facilitated access to files increases the value of a dataset to the scientific community and enhances its impact on population research. The more researchers share data and use it for scientific investigation, the greater the potential for producing important results or insights, and the more expansive the potential impact on public policy. As more outside investigators utilize NIH-funded data in their secondary research projects, support for data collection within the field becomes stronger, as does the justification for future NIH budgets. PIs who have demonstrated that their data continue to be accessed after the original research project is complete are better-suited to receive further research money. When recognition is bestowed on a DBSB-funded investigating team, the Branch is credited with supporting important scientific work. Insum: the more widely a dataset is shared, the more "everyone wins."

[3] http://grants.nih.gov/grants/policy/data_sharing/index.htm.
[4] This includes basic research, clinical studies, surveys, and other types of research.

Archives and Data Sharing

Datasets created by DBSB-funded research vary in their size, complexity, sensitivity, and potential for broad use by the scientific community. This variation has important implications for data-storing and sharing strategies. Some datasets can simply be placed on the World Wide Web and made freely available for downloading. Other datasets are shared "on demand" by providing a copy of a data file to researchers who request use of the data. Often, the PI prepares a public-use data file, which can then be either stored in-house or in an archive.

Why Archive?

The DBSB advises PIs to make arrangements for data archiving for many reasons. Archives reduce the burden on PIs by assuming the responsibility and costs of distributing data and responding to routine requests for data support. There are legal advantages to placing data in archives, as well. High-quality archives, such as the ones described earlier in this paper, conduct thorough reviews of data so that respondent confidentiality problems can be identified and resolved before the data are released to the public. It is a great advantage to have the benefit of this legal expertise before releasing data.

Recent legislation has increased the urgency for DBSB investigators whose data are relevant to public policy to consider archiving their data. It is mandated that PIs release their scientific data on demand through the Freedom of Information Act (FOIA) when the data are: "(1) first produced in a project that is supported in whole or in part with federal

funds, and (2) cited publicly and officially by a federal agency in support of an action that has the force and effect of law[5].” To inform potential grantees about the basic scope of FOIA’s recent amendment, the NIH now requires all Request for Application (RFA) announcements to provide information regarding “public access to research data through the Freedom of Information Act.” In the text included in RFAs, the NIH advises applicants that taking the step of creating a public-use file that meets confidentiality requirements, and putting that file in a public archive will better protect the dataunder this act.

Under the recent amendments, FOIA-based requests for data are sent to the federal agency that supported the research. The FOIA officer of that agency then decides how to provide the data, subject to human subject regulations, and obtains the data from the PIof the study. It is the FOIA officer, not the PI, who determines what data are released. If the PI has archived the requested data, then, under current NIH practice, the FOIA officer informs the requester that they should obtain the data from the archive. Archiving in these cases allows the PI to structure the way data are released in advance of the requests.

Archiving also benefits the secondary users of data. By taking advantage of economies of scale in storing, marketing, and distributing datasets, and by providing technical assistance, archives lower the cost of access per secondary researcher who wishes to use a dataset. Social scientists and others searching for a dataset to help them answer specific research questions save time and effort visiting archives, which increasingly permit free downloads of data from a Web page (where appropriate) and provide Web-based

[5] Identifiable data on human subjects are not available under FOIA, even under the new regulations (45 CFR 74.36).

documentation, search, retrieval, and extraction software, as well as technical assistance and tutorials of varying degrees of complexity. Archived data that are available for use by researchers from a range of backgrounds are more likely to undergo diverse analyses, which can be used to guide policy makers, form opinions, and stimulate further research. Increased use also means increased scrutiny of data. When errors or quirks are discovered by someone working with an archived dataset, this information can be reported to a single source, which can then more efficiently disseminate this information than if it were passed along by word-of-mouth. Archives can also more effectively promote the spread of best practices in data collection and measurement techniques. Secondary researchers accessing archived data can be assured they are working with a "final" version of a dataset, rather than a modified duplicate.

Archive Models, Services, and Costs

To increase the options available to PIs for data storage and dissemination, the NICHD has subsidized the creation of archives through the Small Business Innovative Research (SBIR) program, infrastructure grants, regular research grants and contracts, and interagency agreements. Some of this funding has been allocated to (public or private) "centralized" archives. In a centralized archive of high-quality, many datasets are housed in a single location; all are checked, cleaned, and processed according to common standards and set criteria. Thesecentralized archives often provide centralized control over the conditions of supply and distribution of the data they store. Centralized archives generally provide documentation and access standards that are the same for each dataset, while providing centralized support service and publicity.

One example of a centralized archive is the Inter-University Consortium for Political and Social Research (ICPSR). The ICPSR is primarily supported by member dues and is available to member institutions only. In some cases, a government agency will set up specialized archives at the ICPSR through special funding arrangements for general public use. The ICPSR maintains and provides access to a vast array of social science data for research and instruction. Housed at the University of Michigan, the ICPSR is a membership-based organization, with over 400 member colleges and universities around the world. The ICPSR stores permanent backups of each dataset both on- and off-site. Preserving a version of the data in a format and medium that won't become obsolete gives the ICPSR the necessary flexibility to migrate each dataset to a new format as changes in technology warrant. The availability of data is publicly announced by the archive, and a description of each investigator's study is listed on their Web site. Authorized researchers are able to receive ICPSR files and documentation either directly from the archive or, often, through the Web site of a member institution. In addition, the ICPSR provides technical support, as well as assistance in identifying relevant data for analysis.

Another example of a centralized archive is Sociometrics Corporation, a for-profit research and development firm specializing in social science research applications. In addition to housing data and documentation from over 200 studies, Sociometrics' data archives facilitate data sharing among social scientists. First, Sociometrics adds value to datasets relinquished by PIs by making them easy to use. Once a PI turns over a dataset,

Sociometrics produces extraction software and users' manuals. Archived data are then disseminated in CD-ROM and Web-downloadable formats with Sociometrics' proprietary software that offers search and retrieval, online data analysis, and data extracting/subsetting capabilities. Data users whose analyses require the use of several data sources benefit from archives, such as Sociometrics, that house multiple sources of data at a single site. That a high proportion of customers purchase the complete archive, even though it is possible to purchase access to individual datasets indicates the value of the centralized archive.

Another model for a centralized archive is found at the Henry A. Murray Research Center at Harvard University. The Murray Center data archive is unique in that it accepts not only quantitative data, but also qualitative materials such as case histories, narrative interview data, and audio and videotapes. The Murray Center checks, cleans, and processes the studies it acquires. Use of data archived at the Center is by application only. The Center offers support services and provides access to survey data just as the ICPSR and Sociometrics do. Qualitative materials are also made available for secondary analysis after de-identification. The conditions for use depend on the sensitivity of the data. Users are charged a nominal fee if data are distributed by CD-rom, or extensive copying is required; otherwise the Center's services are free of charge.

For PIs who are uncomfortable with the idea of relinquishing control over the format of their data or over the technical expertise offered, centralized archives may not be the optimal way to share data. Technological developments increasingly allow distributed

models to generate the same benefits as single-location systems. In the typical distributed, or "hub and spoke" model, researchers access data from several different archives through a single "hub" that has agreements with various suppliers and housers of data. Hubs are networked so that common standards and administrative procedures can be maintained, including agreements about the supply and use of data. For member organizations, access to archives is often possible for those affiliated with the organization through their Web site. The University of Virginia, for example, provides access for those affiliated with the university to numerous government and privately collected datasets, including 1990 Census data, U.S. Historical Election Returns, and Sociometrics data, through their Geo-spatial and Statistical Data Center. Many members provide a link to the ICPSR's data through their own institution's Web sites. Columbia University's Electronic Data Service (EDS) archive provides links to the ICPSR archive, the National Center for Health Statistics (NCHS) data warehouse, the federal government's Depository Library Program, and the Roper Center at the University of Connecticut.

Some "hubs" provide their own technical assistance beyond that which is provided through the "spokes." For data users who might not be familiar with these data, EDS staff conduct guided introductions to the archive and provide detailed background and access information for each dataset accessible through their hub. EDS staff are also available to help instructors and researchers find and extract appropriate data. Data users who require technical support can submit e-mail requests or contact data center staff directly. The EDS provides additional services to data users. Because the center "represents

Columbia" in two major organizations of data users, it is able to use these contacts to acquire access to other datasets that Columbia users request. The Harvard-Massachusetts Institute of Technology (MIT) Data Center also offers assistance with purchasing data from other sources byoffering to help negotiate a discount or mediate the data acquisition. Occasionally, the archive will pay a portion of the cost of data acquisitions outside of the ICPSR, if the data is deemed valuable to the wider Harvard and MIT communities.

Though there is a strong sentiment within the population research community that data sharing should be this easy for all types of data, the cost of providing this assistance is significant, particularly for the archives that make live staff available. Currently, the price charged to users for technical support varies across archives. Many archives, such as ICPSR and the Sociometrics Corporation, provide assistance mechanisms at no extra cost with the purchase of membership. The Roper Center in Connecticut also offers consulting services free to members. However, some archives charge fees up to $100 for extraction software[6], while others charge hourly consulting rates to cover the costs of providing technical assistance. Members of the research community have expressed concern that, although these costs may not be a burden on funded researchers, they may be prohibitively high for students and others who don't have outside funding.

Mechanisms for addressing confidentiality

Regardless of the ways in which data are shared, investigators are required to assure that the terms under which participants gave their informed consent, including promises of confidentiality, are strictly maintained. The discussion about how to ensure

confidentiality of shared data proved to be the most participatory, creative, and contentious discussion at the 2001 conference. Though the discussion concluded with more questions than answers, there was broad agreement that the field of population research needs to develop new approaches to preventing inappropriate uses of data, without denying access to data by qualified researchers.

There are many less-than-obvious ways individuals may be identified. Multilevel data organized in hierarchical files linked by identifiers, data that include exact event dates and birth dates, geo-coded information, and qualitative narrative interview data (especially audio or videotapes) all pose confidentiality risks. Longitudinal panel studies are vulnerable to confidentiality breaches during their operational phases, when linkages to identifiers still exist. De-identified datasets may, in some cases, be linkable to other existing datasets that contain personal identifiers. Perhaps least obvious is the risk of deductive disclosure of a research participant's identity resulting from the cumulative information provided by a small set of common (and individually non-identifying) characteristics. Finally, changes over time in regulations or accepted practices for protecting confidentiality can reveal new risks in data archived and distributed under less stringent rules.

Because confidentiality risks and breaches are not always easy to detect, to prevent all kinds of breaches requires a commitment of substantial thought and financial resources to prevention and policing. PIs in the early phases of data preparation can do much to make privacy protection easier down the road. For example, PIs can adopt practices such as

[6] If the researcher who wishes to access the data does not have access to the dataset

never including real names in data and altering date information[7] so that identification is made difficult. Before sharing their data, PIs must take additional steps to minimize the risk that information collected from research participants could be attributed to identifiable individuals. Two strategies are widely used to accomplish this: 1) restricting the data that are shared, and 2) restricting access to the data. The first strategy entails altering the content of datasets or files to be released. Common methods of modifying or "redacting" data to protect confidentiality include deleting cells in tabular data, deleting identifiers, and top coding. In some cases, investigators release only a subset of data, dropping sensitive items or cases. For example, the PIs of the NICHD-funded Add Health dataset, which is housed at the University of North Carolina, make a "public-use" dataset available through Sociometrics. This dataset contains records for only a subset of cases to prevent deductive disclosure, but includes virtually all information collected about each individual in the sub-sample to permit full analysis.

Another way to modify data to protect confidentiality is to use statistical techniques to transform, or "mask" data. Common methods of masking data include substituting simulated data, adding random error, and exchanging values of certain variables between data subjects. Another way to allow statistical access to all variables, while preventing examination of discrete values/cases is to bundle datasets in analytic software packages. An example of this type of interactive tool is PDQ-Explore, which allows users to submit statistical queries and receive tabular and summary results only. Though these techniques are designed to preserve the distributions of variables in a dataset,

[7]Dates can be altered while maintaining information such as duration of exposure, the timing of events in relation to each other, and the historical period in which events occur, etc.

implications of these techniques for data analysis are still controversial, and more research is needed to thoroughly explore their potential.

The second strategy for protecting the confidentiality of shared data is to impose conditions on data access. Strategies for restricting data access include user agreements and contracts, software packages that limit access, and data enclaves.

A simple, commonly used way to prevent data users from making attempts to identify particular individuals, and to keep unauthorized persons from gaining access to the data is through user agreements, which may come in the form of a written contract or an on-screen notification, in the case of distributed computer files. The NCHS has been using an on-screen agreement since 1993, when it began releasing some of its micro data files on CD-ROMs. The user is asked to agree to the following restrictions:

- I will not use nor permit others to use the data in these sets in any way except for statistical reporting and analysis.
- I will not release nor permit others to release the datasets or any part of them to any person who is not a member of this organization, except with the approval of NCHS.
- I will not attempt to link nor permit others to attempt to link the dataset with individually identifiable records from any other NCHS or non-NCHS dataset.
- If the identity of any person or establishment should be discovered inadvertently then no use will be made of this knowledge and the director of NCHS will be advised of this incident, the information that would identify an individual or establishment will be

safeguarded or destroyed as requested by NCHS, and no one else will be informed of the discovered identity.

Many such agreements stipulate that data should be used by certain individuals (university students, faculty, and staff) for certain purposes (academic research) only.

User agreements are difficult to enforce, and violations are punished to varying degrees. The National Center for Education Statistics (NCES) forces researchers to sign affidavits of nondisclosure that make them subject to severe penalties for violation of their oath. To enforce the conditions of their licenses, NCES requires users to agree to be subject to unannounced worksite inspections. Some archives enforce agreements by hiding a unique identifier in each of the datasets, whichallows enforcers to determine the origin of any copy found in unauthorized hands. Most agreements carry financial penalties for violations. In all cases though, user agreements must rely on a certain degree of trust.

Another way of restricting data access to authorized individuals is to create individual copies of analytic software that are matched to individual copies of datasets, so that only the registered owner of the software can analyze the data matched to the software. Because the software provides researchers with the benefit of access to restricted data from the comfort of their offices, these programs not only protect confidentiality, but also make data accessible to communities of researchers very efficiently. Some of this software actually writes an extraction program for a new user and puts it into the format

of one of the available major analysis software packages, thus greatly reducing the time a secondary user has to spend learning how to negotiate the entire dataset.

When contracts, licensing agreements and software are unable to provide adequate safeguards, "data enclaves" provide an option for data sharing. In general, data enclaves grant certain individuals access to a particular dataset on secure computers in a tightly controlled physical setting, then supervise these individuals in their use of the dataset. The Michigan Center of the Demography of Aging (MCDA) is an example of an enclave. MCDA provides approved visitors with office space and high-capacity workstations that allow them to access the statistical analysis software, specialized application software, and utilities necessary to manipulate and analyze restricted data files. The workstations, designed to serve two researchers at a time, communicate with a dedicated server on a network that has no physical connection to any other network or to the Internet. A unique password-protected profile for each user specifies the restricted data files that the user can access based on the user agreement; each workstation allows access to only those restricted data files. Users are allowed to remove statistical analysis results from the enclave only after staff have conducted a disclosure limitation review.

Some enclaves allow only authorized staff to handle and manipulate data. Researchers obtain analytic results through "customized data analysis systems," by submitting program code for their regressions. The code is not only screened for confidentiality issues before being run by enclave staff, but both code and results are also screened before being released to the researchers.

A continuing demand for analyses that require restricted data for small geographic areas such as states, counties, and census tracts, but not confidential identifiers such as names or social security numbers has been the impetus for the creation of the Research Data Center (RDC) located at the NCHSheadquarters in Hyattsville, Maryland. Designed for researchers outside of NCHS, the RDC allows access to data that they would not be permitted to analyze otherwise because of confidentiality/disclosure rules and regulations. Information that would, if accessed with no restrictions whatsoever, be considered identifiable and not releasable can, under the restricted conditions of the RDC, be subject to statistical manipulation. While information concerning named geographic entities cannot be accessed, data ordered by such units can be analyzed at a level not possible with public-use data.

Strict confidentiality protocols require that researchers with approved projects complete their work using the facilities located within the RDC. Prospective users of restricted data must submit a research proposal that is reviewed and approved by a committee, which makes judgments based upon the availability of RDC resources, the mission of the NCHS, general scientific soundness, and the feasibility of projects. It is expected that potential RDC users will develop the research proposal with RDC staff to minimize the time required to complete the analysis. Researchers may supply their own data to be merged with NCHS datasets. These merges are completed by the RDC staff, and the merged files are only made available to the originating researcher, unless explicit written permission is given to allow access to others.

Once proposals are approved, the NCHS provides both onsite and remote access to restricted data through the RDC. Researchers working onsite have the ability to use the full capabilities of the SAS system if they agree to submit to a disclosure review. The analysis capabilities of the remote access system are more limited. In both cases, however, output is thoroughly scanned and screened to ensure adherence to strict minimal disclosure limits; results are suppressed if the minimums are not met. Such procedures prevent intentional, as well as unintentional confidentiality breaches.

Another example of a data enclave is the one housed in the Henry A. Murray Research Center at Harvard University. The Murray Center, in operation for several years, specializes in making very sensitive data available to outside researchers. Such data include original subject records as well as computer-accessible datasets.

Though enclaves provide strong privacy protection, they are undeniably inconvenient, which, several workshop attendees pointed out, imposes a considerable cost on the using public and decreases use. Enclaves can also be enormously expensive to design, create and maintain. To run enclaves such as the ones at MCDA and the University of North Carolina requires a large commitment of highly skilled and highly trusted staff to ensure that all restricted and public-use datasets listed on the user's approved research plan are available on the workstation, and that all security procedures are enforced. Staff hours are also needed to provide assistance with the dataset installation, software installation,

operating system problems, statistical package operation, backups, and user interface issues.

The Add Health study employs several of the aforementioned confidentiality procedures. The Add Health Web site provides a tool for running simple, cross-tabulations based on the public-use dataset (described above and also available through an archive). Access to data not included in the public-use dataset is provided under contracts that require Institutional Review Board approval for the proposed research, and adherence to strict security procedures for handling the data. A small subset of highly sensitive data had been made available to users, only under highly supervised conditions in a "cold room" at the University of North Carolina. However, even highly motivated researchers were unable to use the data under these conditions because of the high costs involved. As a result, specialized contracts were designed to cover offsite use of the data by qualified researchers.

Challenges and Lessons Learned

Although most PIs who receive funding from the DBSB embrace its commitment to provide wide access to data while protecting confidentiality, the DBSB is cognizant that the net social benefits of providing a large number of researchers with access to a particular dataset may be higher than the net private benefits that may accrue to the PIs themselves. Costs to the PI for preparing, documenting, and distributing data for public use can be significant and may continue to accrue for as long as researchers utilize a dataset, a period usually far longer than the funded grant period that produces it. A

significant proportion of the costs associated with data dissemination are due to the expense of providing technical assistance to scientists, especially those who are outside of the normal boundaries of population research, and to students and others who are new to research. As stressed earlier, it is important to the scientific community in general, and to the population research community in particular, that funds be made available to provide technical support to all legitimate researchers, even those outside of the population research field.

When the benefits of providing access are outweighed by the costs of data sharing, data collection projects may not be undertaken, even when they are worthwhile from a societal point of view. Alternatively, PIs who collect data for their own research purposes may refuse to distribute these data to researchers who aren't directly involved in the collection process, so as to retain exclusivity of results. By providing researchers with limited periods of exclusivity, underwriting data sharing costs, and alleviating PI administrative burdens by subsidizing archives, the DBSB increases the benefits of data collection, storage, and dissemination, while decreasing the costs of these activities.

Theoretically, NIH should be able to underwrite the sharing of data collected by PIs. Currently proposed NIH data-sharing guidelines instruct PIs to include the costs of cleaning, storing, disseminating, and protecting data in their proposal budgets. However, in practice, a variety of problems tend to undermine support for data-sharing costs. First, though applicants are asked to address their plans for data sharing in their grant applications, and to include sufficient funds in their budgets to prepare, store, distribute,

and support data, the funds required for data storage and dissemination are difficult to estimate at the beginning of a project and are difficult for reviewers to evaluate. Study section reviewers are usually not themselves specialists in data sharing and are often reluctant to penalize a good project for murky data-sharing plans or inadequate budgets. Furthermore, unexpected increases in data-sharing costs due to changing technologies and changing regulations may catch PIs by surprise, despite careful planning.

The NIH currently has few tools with which to enforce data-sharing standards, relying mostly on a PI's sense of duty to the research community, regard for their reputations, and desire for future financial support to ensure data collectors live up to their promises. The DBSB currently encourages PIs to draw on the experiences of past grantees, to consult with an NIH program administrator, and to consider partnering with an archive early in the process of formulating proposal budgets. Workshop participants widely agreed that the field would also benefit from a more accessible reservoir of expertise about data sharing than currently exists.

Even if PIs are able to perfectly predict the funds necessary for the collection and distribution of their data, adequate funds for data sharing are rarely awarded. Budgets for NICHD grants are negotiated down before awards are made. Large data-collection projects usually experience the biggest budget cuts, despite the large fixed costs of such projects, and the disproportionate impact that such cuts have on sample size. PIs, are faced with the choice of either compromising data collection, or cutting back on data sharing and other expenditures. Often and understandably, the latter solution is chosen.

When projects end, a lack of funding usually ends active data sharing and support by the PI. Breaks in funding often cause interruptions in data sharing as well.

Consideration of past challenges and how they were overcome may help current and future PIs circumvent difficulties. In particular, PIs can learn from the comments of those who represented the NSFH data collection project at the DBSB workshop. The PIs for this study related that, as expensive as state-of-the-art dissemination systems are to maintain, it is more costly to shut down a resource that is widely used by the scientific community. The process by which the NSFH released data to the public had been recognized to be a gold "standard." Two years ago, an unforeseen delay of their grant renewal forced the NSFH to lay off specialized personnel, forcing a temporary shut-down of Web access to their data. The shut-down left several users stranded without data, including an investigator recently funded by the National Institute of Drug Abuse (NIDA)[8]. Though, in the end, considerable pressure from other inconvenienced users resulted in the NICHD providing the support NSFH needed to restart their access capability, the shut-down caused significant damage. Not only was important research delayed, but the additional costs of hiring and training new technical staff to replace those who had been laid off also turned out to be more costly than keeping the dataset up and running.

The challenges faced by the PIs of the Integrated Public Use Micro data Series (IPUMS) were of a different nature. The IPUMS model for making historical population data

[8] His entire project depended on information about the NSFH sample that could only be obtained by the specialized services formerly provided by the user support group.

available for scientific investigation has been copied in a growing number of countries. With support from the NICHD, the IPUMS research team pioneered the technique of combining and reducing vast quantities of historical census data into scientifically useful datasets, and then releasing these data to users free of charge over the Internet. The complete IPUMS consists of 25 samples,which span the censuses from 1850 to 1990, collectively comprising our richest source of quantitative information on long-term changes in the American population. Some of these samples have existed for years, while others were created specifically for this database. Because of the longitudinal nature of the dataset, one would expect it to be possible to study changes over time. Unfortunately, because each sample was created at a different time, using inflexible formatting mechanisms, each year has a different record layout, coding scheme, and documentation. This data incompatibility has complicated efforts to combine the data for longitudinal use. The lesson to be learned from the experiences of the IPUMS is that the process of formatting data must be flexible enough to efficiently incorporate changes in standards that evolve along with technological changes[9].

Future Challenges

The current state-of-the-art of data archiving and access provides a range of options for data collectors, archives, and users. These options are efficient because they increase the likelihood that PIs will be able to design data-sharing strategies that are appropriate to their datasets. For this reason, though standards must be put in place to ensure the quality

[9] This dataset, too, experienced funding shortfalls. Had it not received infrastructure funding from the DBSB and the National Science Foundation for a time period beyond the life of its initial grants or partnered with ICPSR at Michigan, the data might not be as accessible as it is today.

and confidentiality of shared data, the solution is not to force all investigators and archives to employ the same means to achieve such quality and confidentiality.

Although it is clear that problems continue within the current system of data storage and dissemination, the future promises to bring even more challenges. Technological advances in the biological, computer, and social sciences have made it increasingly possible for researchers in all fields to acquire, interpret, and disseminate data that were previously unattainable. For example, information gathered by swabbing a person's cheek can reveal elements of that individual's genetic makeup. Innovative survey methodologies allow PIs to collect unprecedented information about the social networks of teenage study participants. By combining designs and methodologies, researchers can append ethnographic and observational information to quantitative survey information collected about research subjects. Several large-scale projects that are currently underway use computer algorithms to match census data about individuals to data collected about their place of employment, work, and salary history. Health care agencies and organizations collect extremely detailed information about the health status of individuals. As exciting as these developments are for researchers seeking to uncover knowledge that may lead to improvements in the health and well-being of individuals, they are also frightening to those concerned about the privacy of individuals about whom these data have been collected.

DBSB grantees have been at the forefront of the hunt for improved archiving methods. As the threat of future legislative action or litigation places greater amounts of data at risk for FOIA disclosure, the need for innovative approaches becomes even more critical. Although future technological changes may improve the efficiency of data distribution systems, therefore reducing the cost, the fact that researchers can collect more and more sensitive data, combined with the current and evolving atmosphere of heightened concern about the protection of human subjects and their privacy, may cause data dissemination systems to become more expensive. There is an urgent need to ensure that funds will be available. New funding is needed because the current sources of support for dealing with data sharing, which have come through individual research grants for data collection and from the indirect support of population research centers, are on the verge of being overwhelmed. In the past, these sources have provided funding for the design, implementation, and sharing of data from large complex projects[10]. However, center directors are clearly feeling overwhelmed by the impending financial stresses associated with these projects.

Panel Recommendations

What follows is a summary of the recommendations of the data workshop participants, followed by a discussion of how the NICHD might address each of the issues.

[10] The PIs of Add Health and NSFH readily stated that they could not have even attempted these efforts without center support.

1. **All reasonable steps should be taken to ensure that data collected under NIH assistance mechanisms are made accessible to all who wish to use them for scientific analyses.**

Assuming that demographic researchers are willing to adhere to the standards established by their field, and to provide appropriate access to their data after a reasonably short period of exclusivity, there are two main reasons for PIs to restrict access to their data: privacy concerns and data sharing costs. New ideas and technological advances are needed to effectively address the first barrier. Additional funding and the development of more efficient methods will eliminate the second. One "new idea" that was suggested at the workshop was the possibilitly of turning one's office computer into an enclave using software or remote monitoring devices. Though "front-end software" systems are a more secure and flexible alternative to policing by contracts, the software solution, as it currently exists, is inferior to the data enclave solution because it is not possible to review the data output of users who might be able to circumvent even the most advanced confidentiality systems. This screening capability may be added in the future if, for example, it becomes possible to remotely monitor output from a secure site.

To increase access to data for inexperienced and seasoned researchers alike, PIs must make technical assistance available to users. This recommendation is particularly relevant to projects that yield major datasets. Though charging by the hour for support, or turning data over to an archive that charges a service fee is a viable option, participants expressed concern that these solutions would increase access only for those who could

afford to pay. Subsidizing data access for unfunded researchers could prevent certain researchers from being priced-out of using the data, including pre- and post-doctoral students who are just starting a line of research that may last a full career, and researchers from allied fields who may wish to use the data for a limited number of analyses. Panelists recommended that PIs be encouraged to include the cost of subsidies and technical assistance in their budgets when submitting NICHD proposals for funding, and that the NICHD either be more aggressive in keeping technical assistance costs in the project, or fund supplemental requests later in the course of a project when dissemination costs are better understood. To reduce the burden of uncertainty about whether actual costs of providing technical assistance will exceed budgeted funds, PIs can partner early on with an archive that forces users to pay for technical assistance. Another proposal to ease burden on PIs and their host institutions *ex post* would limit the amount and type of technical assistance allotted to data users. This option is clearly a second-best solution because implications are that, in the future, fewer people will have access to publicly funded datasets.

2. **The DBSB should increase its support of data sharing by investing new funding in improving the data-sharing infrastructure of the population research community.**

Panel members agreed that increasing support for data sharing should be the foremost goal of the DBSB. This goal might be achieved by means of assistance mechanisms **that** support three types of infrastructure, as suggested by workshop participants:

infrastructure to assist PIs in the planning and designing of data sharing for complex projects; infrastructure to simplify implementation of data access while the project is underway; and infrastructure to maintain access to data after the project producing the data has ended. To provide these infrastructures, the conference participants generally agreed thatnew funding is needed.

The fact that the technology for dealing with data sharing is still evolving implies that infrastructures put in place now should be flexible enough to incorporate new technology as it is developed. Because a single, omnibus solution to all of the problems involving data sharing is neither feasible, nor desirable, conference members suggested that multiple approaches be woven into a loose, but coherent overall strategy that draws on the capabilities of multiple organizations. Many center directors are looking to centralize aspects of data storage as a way of avoiding duplication and of cutting costs, while retaining individual control over distribution and access. Because doing this requires more flexibility than the DBSB centers program has traditionally provided, the Branch has transformed its centers mechanism to make it a more flexible program. The conference participants suggested that new infrastructure grants be used to provide, for example, advice for designing and implementing data sharing, specialized enclave facilities, and long-term data archiving after individual project support ends.

3. **PIs should be encouraged to fully articulate and budget for the funds to pay for the costs of cleaning, documentation, storage, archiving, and distributing their data.**

Panelists agreed that one of the best ways to improve both the efficiency and the effectiveness of the current data-sharing state-of-the-art is to encourage PIs to think about archiving their scientific data from the earliest possible point in their research. The NIH currently requests that PIs include a description of their archiving plan in the study design and projected archiving costs in the budget of their applications. In addition, the NIH is asking applicants to re-think the typical informed consent and human subjects procedures so as to reflect the wide range of potential uses for the data collected. Planning ahead is important not only to ensure that data are collected and formatted in a way that will minimize back-end formatting when the data are archived, but also to ensure that the funds necessary to adequately archive and disseminate data are available at the end of the collection process.

4. The DBSB should use assistance awards as part of its commitment to avoiding cuts to/discontinuities in funding used for data-sharing.

Because unexpected cost overruns, budget cuts, and advances in the field of data access and archiving make it difficult to predict whether budgeted funds will be sufficient to accomplish all of a researcher's study goals, the NICHD was urged to be vigilant in protecting funding for data sharing during the course of data collection projects. Protecting data collection projects that have significant data-sharing costs from large budget reductions is an important first step. Providing emergency, administrative supplements to researchers who find themselves with insufficient data-sharing

fundswould also help to reduce the impact of uncontrollable and unpredictable elements in the data collection process.

Current proposed guidelines for data-sharing will focus the explicit attention of review groups on data-sharing needs, and may stimulate many requests from active grants to provide either administrative or competitive supplements to enhance data sharing. The NICHD, as well as all other NIH Institutes, Centers, and Divisions, should prepare to deal with these requests.

5. The DBSB should help to identify "best practices" for documenting, storing, distributing, and archiving scientific data in the population sciences, and to make this information available to inexperienced PIs who are formulating archiving plans.

Panel members suggested that the overall process of data sharing could be made more efficient if the DBSB invested in facilitating the refinement and dissemination of standards, practices, and procedures ("best practices") for archiving and accessing scientific data. The DBSB was urged to take leadership in convening conferences designed to identify such "best practices", where PIs and archives could learn from each other about what does and doesn't work efficiently. Information about the costs of different archiving options would be a big help to first-time data collectors, while attendance by representatives from archives might help these groups drum up business from those unfamiliar with the competition among archives. Through the Population Research Infrastructure Program, the DBSB could allow one or more institutions to develop the

capacity to assist investigators of population research data collections in employing the field's the "best practices" in data sharing. Though many participants cautioned that the Branch should take care not to adopt rigid standards in a realm where technology is evolving at a rapid rate, it was generally agreed that much could be done now that would aid the entire field.

6. The DBSB should encourage companies that have the capability to develop innovative data archiving technologies.

The magnitude of the societal benefits associated with lowering the cost of access to secondary data ensures that sharing technology will continue to advance. One way to encourage innovation in the area of privacy protection would be for the NICHD to provide more SBIR and STTR funding for technological innovations in data sharing. The DBSB funds could also set up competition between public and private archivers to advance the technology of the archiving processes and find a more efficient, less costly way to disseminate data. Alternatively, the DBSB could provide grants to companies that have the capability to develop innovative data archiving technologies to move the field along.

We have summarized a variety of archiving and access models in this paper. To determine which method is best for a particular dataset requires an understanding of the underlying structure of the survey, data, and links to other data. It is clear that, because

of the complexity of modern population research datasets, it is inconceivable that the primary responsibility for creating access to scientific data should be vested in anyone but the PI and the investigating team. Though he or she may relinquish control to an archive, there is no one better than the PI to make data storing and access requirements explicit. However, even experienced PIs are often naïve about the complexity of data sharing in the beginning stages of planning a large study. Indeed, at this conference, even the most revered pioneers of data sharing in population research admitted to being generally unprepared for the dissemination demands made on them. In fact, though all present appreciated that it was important to address these issues from the beginning, no one at the conference claimed to have been able to fully think through all of their data-sharing issues in advance of actually doing it.

Time is of the essence. For many years, the DBSB has relied on the resources and creativity of its PIs and research institutions to solve the problems attendant to data sharing. The business of data sharing in population research has gotten very complicated, and the system of data sharing within our research community is in need of help to ease the burdens on centers and PIs. The recommendations presented here are sensible steps that may enable the population research community to maintain the degree of data-sharing to which the field to which has grown accustomed and to allow the research community to continue to innovate so that the promise of a rapidly evolving research frontier can be realized.