Expert Meeting on Guidance on Data Sharing for NIA Studies

The National Academies of Sciences, Engineering, and Medicine
Committee on National Statistics

May 10, 2021

Web Conference

September 27, 2021

This meeting summary was prepared by Rose Li and Associates, Inc., under contract to the National Institute on Aging (NIA). The views expressed in this document reflect both individual and collective opinions of the meeting participants and not necessarily those of NIA. Review of earlier versions of this meeting summary by the following individuals is gratefully acknowledged: Christina Tricou, Alison Aughinbaugh, Dana Carluccio, Noura Insolera, David S. Johnson, Kathryn Lavender, Rose Maria Li, James McNally, Alita Nandi, John Phillips, Lucas Smalldon, Ana Trisovic, Nancy Tuvesson, Lars Vilhuber.

Table of Contents

Meeting Summary

Introduction

The National Institute on Aging (NIA) commissioned the Committee on National Statistics (CNSTAT) to convene an Expert Meeting to consider best practices for sharing longitudinal data sets. The meeting was held via web conference on May 10, 2021, in conjunction with CNSTAT's spring annual meeting.

Because longitudinal datasets have a high volume of data and complex links within individuals and families among waves of data collection, they pose unique challenges to making data public. The meeting focused on five major topics: (1) providing documentation and easy access to and use of datasets, (2) developing interfaces for accessing and finding variables, (3) creating longitudinally linked files, (4) training new users, and (5) maintaining a repository of user-generated code and data.

Representatives from the study teams of large longitudinal surveys (i.e., National Longitudinal Survey of Youth, Understanding Society, and Panel Study of Income Dynamics) presented policies and strategies that have best served their teams' goals. Representatives from databases and projects that focus on teaching data sharing best practices to researchers (i.e., National Archive of Computerized Data on Aging, American Economic Association Code and Data Repository, and Dataverse) presented on the critical elements of best practices for data sharing and the level and type of support data providers and users need to meet those best practices.

From these presentations and ensuing discussions, the Expert Meeting identified a set of best practices that can guide future data sharing endeavors. The meeting agenda and list of participants are included as Appendices 1 and 2, respectively.

Facilitating Greater Use of NIA Longitudinal Studies: Data Documentation

Mick Couper, University of Michigan

National Longitudinal Survey of Youth

Alison Aughinbaugh, Bureau of Labor Statistics

The Bureau of Labor Statistics (BLS) conducts and publicly releases data from two National Longitudinal Surveys (NLS), making BLS staff who support these surveys experts on the challenges of sharing longitudinal data. The National Longitudinal Survey of Youth 1979 (NLSY79) began in 1979 and enrolled participants born between 1957 and 1964. The National Longitudinal Survey of Youth 1997 (NLSY97) began in 1997 and enrolled participants born between 1980 and 1984. NLSY97 presents more complications for publication of data because its design was more complex from its launch; NLSY79 began with paper-and-pencil collection, necessitating greater simplicity. NLSY97's more complex design involves different participants

within each wave answering different questions, making both harmonization and comparisons across and within individuals more challenging than if all surveyed individuals answered the same sets of questions in each wave.

Providing Documentation and Easy Data Access and Use

In its public data releases, BLS provides three types of data files: (1) a public use dataset, which contains variables measured as well as weights assigned and includes one geographic variable indicating region in which the respondent lives; (2) a confidential geocode dataset, which contains state and county identifiers as well as information such as colleges attended; and (3) a restricted-access dataset, which contains participant ZIP codes and census tracts. The restricted-access dataset can be accessed only at the BLS National Office or in a U.S. Census Bureau Federal Statistical Research Data Center (FSRDC) after applying to BLS. BLS provides extensive online documentation to accompany NLS data both at the BLS website and NLSinfo.

Developing Interfaces for Accessing and Finding Variables

An online searchable bibliography enables data searches by topic or cohort. Each cohort has a dedicated page on NLSinfo that displays all documentation, including a topical guide, information about the sample, explanations about how to use and understand the data, and all associated questionnaires. The topical guide divides data into 12 categories that are consistent across cohorts and provides documentation for each category, consisting of explanations of how variables within the categories were collected and how they varied over time. BLS further subdivides categories and lists created variables relevant to those categories at the top of each category's webpage to inform users of their construction.

All public data from NLSY79 and NLSY97, as well as codebooks (describing the contents, structure, and layout of data collection so that all variables in a dataset are explained), are available through NLS Investigator, a tool that enables users to navigate data in various ways. Users can browse an index listing of all collected variables by category, explore dropdown menus, and search by variable type, survey year, and keyword. After finding the variables of choice and tagging them, users can download their tag sets in SAS, Stata, SPSS, or R format along with a codebook for those specific variables. Users can also find explanations of how the created variables are developed, as well as raw components of those variables to replicate variable creation.

Creating Longitudinally Linked Files

Data harmonization across cohorts has presented a challenge for BLS. NLS Investigator provides a beta version of a harmonized data set using only status variables, such as highest degree completed and marital status, but even these simple measures were difficult to harmonize because of differences in how they were recorded. The next step for BLS is to develop documentation describing how different variables were defined and were measured across cohorts, facilitating harmonization across datasets in which variables are not perfectly aligned.

Training New Users

Most training on use of NLS datasets is conducted between academics who use NLS data and their graduate students. However, BLS has run data workshops to train users on ways to best

approach the data and provides annual meetings of the Population Association of America (PAA) and, through 2008, week-long summer workshops funded by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD). BLS also provides online tutorials and new user tabs on its website, as well as responsive user services that address user comments, concerns, and questions.

Maintaining a Repository of User-Generated Code and Variables

Although BLS provides code for constructed variables, it does not provide a repository for user-generated code because of concerns that this code would be considered a BLS product, making BLS legally responsible for its suitability.

Understanding Society

Alita Nandi, University of Essex, UK

Understanding Society, also known as the U.K. Household Longitudinal Study (UKHLS), is a survey of a nationally representative sample of U.K. households interviewed annually that has successfully and effectively conveyed its data to the public. Participants from the British Household Panel Survey, which began in 1991, were added to the UKHLS in 2010, the year after its launch. Data from both surveys have been harmonized and released since 2016. The survey is administered separately to two categories of respondents: (1) people aged 16 and older (i.e., "adults") and (2) people aged 10-15 years. Although the survey includes information on children under 10, they are not directly surveyed and information is collected only from parents. The survey covers many topics, with a core set of questions asked annually and rotating modules implemented every 3, 4, or 6 years. The study collects some retrospective information (e.g., past employment, fertility, and partnership are collected at the initial interview). The large sample size and immigrant and ethnic minority boost samples make this study the only longitudinal data resource for research on immigrants and ethnic minorities. In 2012-2013, nurses visited sample member's homes once, to collect health and biomarkers. Since April 2020, the study invited its 16+ sample members to complete a short 20-minute online survey (monthly until July, then bi-monthly) about their experiences during the COVID-19 pandemic including health and serology data.

Providing Documentation and Easy Data Access and Use

UKHLS's online documentation, resources, and trainings are designed to provide data users with multiple pathways to the same information in a user-friendly manner. The study staff make data accessible not only to academic researchers, but also to policy experts, journalists, and other lay researchers, even those unfamiliar with longitudinal datasets. To that end, the website explicitly states that data are available wave-by-wave.

Users can find more than 300 datafiles covering 60,000 variables on UKHLS's website. Data from each questionnaire, or instrument, are provided in separate fields, and each wave has a different file with the same root name and a wave prefix. A cross-wave file includes time invariant information, such as a participant's birth year. Until 2019, the user guide was available only as a large PDF; for ease of use, the study team designed an online modular version of the

user guide as another option. Additional, topic-specific user guides are available, enabling easy exploration of diverse topics such as Family & Households, Ethnicity & Immigration, COVID-19, and Education.

Developing Interfaces for Accessing and Finding Variables

A tool enables searching both by keyword and by specific index forms, terms, and waves for users familiar with the data structure. The study team intends to add subcategories to keywords. Users who prefer to search by questionnaire modules can search across waves and, within a given wave, can see the breakdown of the questionnaire into modules.

When users retrieve a variable, the system teaches them about the data structure by providing a variety of information on that variable, including in which waves and files it appears, a description of the variable, the exact question text, the eligible participants, and the variable's frequency of appearance in each wave. When accessing a derived variable, users view the variable itself, as well as how it was derived. Users can download any variable of interest. The questionnaires are available to the public, along with long-term content plans that outline which modules will appear in which years.

Creating Longitudinally Linked Files

Because the UKHLS study team wants to share not only the data, but also the process behind developing the data, it provides code that merges files for specific types of research available as simple syntax for various programming languages, including STATA, SAS, SPSS, and R. The study team has started to make the code for creating derived variables, as well as brief notes describing the process, available online.

Training New Users

To further support new users and aid the public, UKHLS provides short YouTube videos focused on different aspects of its data and survey, as well as webinars from different subject matter experts, conference speeches on the survey, and animations that explain longitudinal data. The study team provides three workshops each year with lectures, data management exercise walkthroughs, and worksheets, with all course material available via Moodle. The team also provides half-day weighting and panel data workshops (to be converted into a longer course in 2022), as well as teaching datasets for lecturers. User support services respond to user inquiries via email, one-on-one online chats, or a user support forum, and provide an FAQ.

Maintaining a Repository of User-Generated Code and Variables

UKHLS provides a repository of code and variables generated by third-party researchers alongside contact information for the creator of each syntax file with the goal that users will determine how to manage the data themselves.

All publications that used the data are searchable on the UKHLS website. Finally, to ensure that researchers cite the dataset correctly, UKHLS provides an appropriate citation format.

Panel Study of Income Dynamics

David S. Johnson and Noura Insolera, University of Michigan

The Panel Study of Income Dynamics (PSID) has a genealogic design (multiple generations in different households are included) with a history of effective data sharing. The study team conducted annual interviews from 1968 to 1997 and biennial interviews thereafter. Upon becoming adults and forming their own households, participants' children are invited to join the study, and the study now includes approximately 10,000 families in each wave. In total, more than 82,000 individuals have participated. More than 3,000 individuals have participated continuously since 1968, while others have only participated in a single wave.

Providing Documentation and Easy Data Access and Use

The study team aims to facilitate researchers' use of its study data. To that end, PSID's website encourages users to review the data analysis of publications that have used these data before downloading any datasets of their own. Its library of publications, dissertations, reports, and working papers is updated quarterly. The website also provides codebooks, user guides, variable descriptions, context (including the reasons for asking given questions), and other supporting documentation. The information for each variable helps users understand the data they are working with; for example, the additional context may show the years in which the survey allowed participants to mark more than one answer, for example, for racial identification. All questionnaires and documents back to 1968 are available as searchable PDFs, including those that are handwritten or produced by a typewriter. For each data wave, a user guide describes links to references, which questions were added or removed, and other upkeep notes.

The website hosts not only the main study files, but also ongoing supplements such as those for child development and the transition into adulthood, defined as the period between ages 18 and 28 years. Young adults are invited to join the main PSID study as their own heads of household.

Developing Interfaces for Accessing and Finding Variables

The Online Data Center (ODC) curates the datasets into manageable forms for users. Any user can search and browse the data, but to download the data, a user must register with the study. The ODC automatically merges data across waves and supplements at the individual or family level. Users regularly complete searches across all waves, and receive a custom codebook just for their chosen variables in the statistical programming language of their choice when using the ODC. This custom codebook functionality enables users to avoid working with a documentation page that details every datafile for every family across all 95,000 variables. Any publicly available data can be downloaded through the ODC in this bespoke manner. Carts with full family files and selected variables can be made public so that students or collaborators can easily find the same information. Users can also download full data files and conduct merges themselves, a strategy typically used by researchers who have long worked with the survey and have written the necessary code. Restricted data cannot be downloaded from the website and must be analyzed within a virtual data enclave.

Creating Longitudinally Linked Files

The genealogic nature of the sample enables intra- and intergenerational comparisons, and the Family Identification Mapping System (FIMS) links different family members without prescribing family definitions, allowing people to define family relationships themselves; for example, if a researcher is only interested in full siblings rather than half siblings. FIMS is especially useful to researchers because more than 1,000 individuals who are a current reference person or spouse partner of their own household had a great-grandparent who was ever included in the sample.

Training New Users

Dr. Insolera leads the education and outreach team and teaches a user workshop that hosts 25 students each year. The workshop structure includes both lectures and hands-on labs, and students will emerge from the weeklong workshop with a usable dataset suitable for studying their research question. For those users who cannot attend organized workshops, PSID's website includes a "Getting Started" page modeled after that of UKHLS. Eighteen video tutorials explain a variety of topics and help users begin their work. The user help desk curates an FAQ page from the actual most frequently asked questions.

Maintaining a Repository of User-Generated Code and Variables

PSID also hosts a repository of user-generated code via OpenICPSR (Open Interuniversity Consortium for Political and Social Research) at the University of Michigan to enable community sharing.

The National Archive of Computerized Data on Aging (NACDA)

Kathryn Lavender

NACDA, located within ICPSR, has extensive experience with studies that have used both effective and suboptimal data sharing practices. It has hosted data for secondary use for more than 35 years, with data from more than 1,500 studies, 500 of which are longitudinal. The data pipeline approach used by NACDA and ICPSR makes this volume of data manageable and enables the archive to host not only complete datasets but also datasets from ongoing studies.

NACDA's goal as a repository is to guide investigators in the use of best practices for data repositories to facilitate ease in data use and sharing. These best practices include the following:

1. Addressing the risk of respondent confidentiality,
2. Providing complete collections of metadata,
3. Providing complete variable metadata, either directly in the datafiles or in syntax files,
4. Providing documentation, such as codebooks, questionnaires, and user guides, and
5. Especially for longitudinal data, the provision of complete information packages that include explanations of how to use weights and how sampling occurred, as well as questionnaires and any assistive materials, ensuring that data users can utilize data to the best of their ability and see full research potential.

Providing Documentation and Easy Data Access and Use

Searches of NACDA's database generate results at the study and variable level as well as across data-related publications and the website's non-study pages (e.g., announcements). Documentation is viewable without logging in and can be previewed before it is downloaded. Documentation always accompanies downloaded data. NACDA uses the Data Documentation Initiative (DDI) Codebook, an XML markup language that makes it easy for users to understand the data they are viewing and therefore select and compare variables. The NACDA website provides study datafiles in a variety of formats (e.g., SPSS, SAS, STATA, R, ASCII, and CSV).

Developing Interfaces for Accessing and Finding Variables

In addition to ICPSR, NACDA works with Colectica, a web application for publishing data using open data standards, to facilitate efficient comparisons of longitudinal aging data. Datasets in the portal are findable by name, question source, and concept. Users can select certain variables to create customized data and documentation files.

Creating Longitudinally Linked Files

NACDA encourages data providers with longitudinally linked files to supply the syntax and code required to merge or link files, as well as examples of how to extract subsets. PSID, UKHLS, and NLS all follow NACDA's guidelines.

Training New Users

To train new data users and providers, NACDA provides multiple resources on data sharing best practices, including webinars and both email and e-visit contact options. Secondary data users also have access to a YouTube playlist, podcasts, and ICPSR summer programs for further training.

The most important means of helping secondary data users is ensuring that data providers share all relevant data from the outset. In pursuit of that goal, the NACDA website's data deposit page features the data sharing best practices and a checklist, and encourages providers to use consistent file naming conventions, provide complete metadata, and provide syntax for merging files if necessary.

NACDA also works to ensure reproducibility of its hosted datasets, by (1) applying consistent standards at the file, variable, study, and series levels, (2) considering the accessibility of data products, (3) working with data providers to obtain complete and quality materials, and (4) utilizing feedback from data users and the research community when possible.

The primary challenges that NACDA has encountered and foresees as future concerns include the increasing demand to provide data in short timeframes and a persistent lack of understanding within the research community about data sharing best practices. An ongoing and consistent educational outreach program would help to address these concerns. In addition, NACDA hopes to expand its cross-series comparison efforts through collaboration within ICPSR and other data projects, such as Midlife in the United States (MIDUS) longitudinal project. Finally, a catalogue of Common Data Elements would ease the longitudinal data sharing process by developing consistency across datasets.

Discussion

Excessive Variable Provision

The panel discussed how to determine the point at which the number of variables in a dataset overwhelm users or pose storage challenges, as well as the percentage of data provided that is actually used. Via the ODC, PSID staff can identify the variables downloaded and the keywords used, signaling users' intentions when downloading the full dataset. NLS also tracks how often each variable is downloaded and reviews that information annually. This information helps NLS to determine whether a given variable is not useful and can be removed from a survey.

Users' Most Frequent Requests

The panel discussed the problems that data users tend to encounter and how data providers can address those problems. Often, users pose analytic questions about which test or model would best serve their research, which only large educational outreach programs—and not the data providers—can answer. Inconsistent variable names pose challenges for users, which would be alleviated by maintaining consistency across years. The panel praised UKHLS's FAQs, which clearly delineate the limitations of the data and describe the pros and cons of downloading the entire datafile. Finally, the panel discussed the potential benefits of conducting a cross-study examination to determine the reasons underlying study teams' differing data presentation and sharing choices; such an examination could improve approaches from all data providers.

Addressing User Needs with Complex Data Systems and Multiple Data Sources

Elizabeth Stuart, Johns Hopkins University

American Economic Association Code and Data Repository

Lars Vilhuber, Cornell University

Based on his work with the American Economic Association (AEA) journals, Dr. Vilhuber has formulated three primary goals to improve data sharing: (1) helping researchers to improve their practices, (2) helping journals and associations to assist authors, and (3) helping data providers assist authors. The primary challenge Dr. Vilhuber has encountered in his attempts to improve data sharing is the lack of author documentation and appropriate citation in datasets, which obscures data provenance. Dr. Vilhuber helped develop a Data and Code Availability policy that requires authors to appropriately cite, document, and share data and code whenever possible. That policy was implemented by AEA in 2019, institutionalizing de facto processes that had been adopted since 2004. The information provided by authors should be sufficient to replicate their findings and is thus called a "replication package."

To support the Data and Code Availability Policy, Dr. Vilhuber's team conducts narrow computational reproducibility checks before publication of a study in AEA journals and requests author revisions where necessary. Submitted studies often fail these checks because of a failure of curation and poor coding and citation practices. Most manuscripts accepted into the system

require two rounds of corrections. Occasionally, the team must contract with third parties with access to data that are either impossible or time-consuming to obtain (e.g., U.S. tax data). For publications, the literature attached to a data package will address some of the same content as codebooks, variable levels, and other results of the ICPSR curation process. Therefore, the team is somewhat lenient on the requirements to share full data documentation.

Importantly, not all data can be shared by authors because of confidentiality concerns. Therefore, authors should explain where and how they obtained their data and should identify who else can access the data. To confirm authors' descriptions of how to access data, Dr. Vilhuber's team preemptively checks and improves code archives even when sharing permissions bars inclusion of those archives in the replication package—provided that the team can access the data itself. The team has executed code using confidential raw data. The team has improved archives and citations, and Dr. Vilhuber encourages data providers to share citations that can simply be copied and pasted. Finally, he encourages authors to share code when appropriate (which is almost always), share data when possible, and display information about the data and code.

Dr. Vilhuber listed other considerations related to data sharing. First, journals must confirm that authors have obtained the rights to share data, because authors are not always aware of the rights or restrictions to which they agreed. Second, authors are not trained to deposit data—a problem that could be alleviated by data providers' development of step-by-step guides to making data searchable and FAIR (Findable, Accessible, Interoperable, and Reusable). Finally, user-restricted files, a lack of Digital Object Identifiers (DOIs) assigned to datasets, and a lack of data-provider-suggested citations complicate researchers' attempts to improve their data sharing practices.

Dataverse

Ana Trisovic, Harvard University

The Dataverse project is a free and open-source software platform intended to improve research data sharing and citing. It provides a repository software usually installed at institutions, but it also supports some entire countries' research communities (e.g., Norway and the Netherlands). Users can access Dataverse with either standalone or institutional accounts and can upload data either with the website's user interface, Application Programming Interfaces (APIs), command line, or any of six software clients written in the following languages: Java, R, Javascript, Ruby, and two in Python. Users can search for variables in the repository that are associated with metadata from tabular files. Dataverse provides extensive guidelines and documentation. Each local installation may also have specific support and documentation tools. YouTube tutorials also assist users in learning how to approach the platform.

Using the Dataverse platform, researchers can access collections of datasets or secondary collections created by individuals, institutions, or journals. Within these collections is a subset of data or replication datasets, often containing a bundle of data, code, metadata, and other

files needed to reproduce a particular published study. All datasets at Dataverse have some metadata, some containing information such as related publications, dataset metrics, supplementary files, and badges from the Center for Open Science; Dataverse repositories provide support for multiple metadata standards in both human- and machine-readable formats. Upon upload to Dataverse, files are converted to tabular formats and descriptive statistics are automatically computed.

By applying consistent standards across datasets, Dataverse intends to improve reproducibility of analysis and reusability of data. Alongside collaborators, Dr. Trisovic conducted a large-scale analysis to assess the reproducibility of R and Python Code from the Harvard Dataverse Repository. Even after code cleaning (such as fixing file paths and ensuring pre-installation of used libraries), only 40 percent of the code in R was re-executable. The most frequent errors were library-based, but missing files, syntax errors, and other issues also contributed. Journals with stricter data policies show higher rates of executable code, suggesting that this problem is solvable. Various web-based cloud tools that capture code dependencies have emerged—such as Code Ocean, Whole Tale, and Binder—which can alert researchers to potential coding problems. These tools replicate a full development environment and are even able to recreate plots. In addition, the Harvard Dataverse Repository interacts with Jupyter Binder, so that if a DOI from the repository is entered into the Jupyter Binder, it automatically imports the associated data and may be able to re-execute code (if present) directly from the web browser.

Open questions and directions for future research and development include conducting research to examine common practices, identifying shortcomings of existing approaches, developing new software, and supporting dissemination of emerging computational components, tools, and infrastructure. Software metadata and DDI-Cross Domain Integration also provide pathways for improvement.

Discussion

Financial Dimensions of Reproducibility and Reusability Efforts

NIA is focused on data reuse, of which reproducibility is one important component. However, the resources required to fund improvement efforts, such as those of Dataverse and the AEA journals, can be expensive in both human effort and finances. The AEA's effort, as detailed by Dr. Vilhuber, is funded by the Association, but that funding may not be sustainable. However, the most important and least costly step will be to train the next generation of scientists to use best practices.

The panel noted that NIH's resource sharing requirements depend on the size of the grant, and NIH Institutes should implement an expectation that data producers will produce and share replicable data. The panel suggested that researchers be expressly paid to develop shared data; for example, funders could withhold 10 percent of a grant until the associated data are public and reproducible.

Responses to Replication

The panel agreed that reproducibility and replication checks must be done early and often, but improving code checking and other data sharing practices remains a slow process. Most journals are participating or will participate in data sharing projects, and authors will learn to adhere to expectations set by editors. However, amid their many tasks, researchers may deprioritize learning data sharing best practices.

Necessary Trainings

The panel agreed that researchers must be trained to value reproducibility, capture all analysis, and provide open sources. Funders, universities, journal editors, and peers can all contribute to improved data training. *JOSS*, an open-source journal for software, enables reviewers to assess public code, which provides critical feedback. The quantity of code available has also increased over time, and, as with *JOSS*, that availability provides useful examples for other researchers. The panel noted that lack of confidence around code may contribute to scientists' reticence to share data. To counter this obstacle, journals can provide walkthroughs and basic coding advice. For example, researchers should not hardcode directories (i.e., they should not embed parameters within underlying programs), but many do.

Data Security

As discussed in many presentations, data security and confidentiality pose challenges to data sharing. For example, while Jupyter Binder is a useful tool for many researchers, it is only intended for use with public data and does not provide sufficient security restrictions for confidential data. Providers with confidential and restricted data must provide resources for computing; for example, the Census Bureau asks researchers to bring code to the data rather than bring the data to the code. To test code reproducibility when data are restricted, some data providers produce simulated datasets. The panel noted that reproducible and open-source are not synonymous, and restricted datasets can still be verified as reproducible through third-party reproduction services.

Version Identifiers

The panel noted that version numbers in datasets are not always sufficient to communicate the nature of the data. Therefore, data providers should use additional identifiers, clarify the potential issues of working with different versions, and provide guidelines for finding and referencing different or older versions. Data are occasionally merged across version numbers, producing another set of reproducibility challenges. Requiring authors to identify the version they used—and training them to keep raw files in a raw request folder—are critical needs.

Wrap-up and Final Discussion and Priorities for NIA to consider

Missing Tools

Curated research data and code are critical for improving reproducibility and scientific reuse. However, the panel noted that massive datasets and variables with different structures and structures that change over time complicate this solution. In addition, data providers cannot help researchers make analytic decisions in pursuit of specific hypotheses or questions. To

reduce the number of researchers requesting such support, providers should design the best possible guidelines for working with data and when contacting helpdesks is appropriate. The panel also suggested that data providers can develop and publicize the types of analysis used by most of the user base.

Credit and Data Citation

The panel discussed the issue of citing and giving credit to the people who produced the data and who generated secondary analytic datasets. Hosting third-party data and code poses challenges, because if projects are seen as responsible for all code they host, they cannot host any code that they have not thoroughly assessed. Whether projects do in fact bear this responsibility is judged differently by different legal experts. Data providers should include appropriate citation strategies whenever possible, which will improve appropriate attribution.

Weighting

The panel discussed the function of weights in longitudinal samples and data providers' approach to weighting the data or providing training or guidance on how to weight the data. NLS has a custom weighting system to enforce its nationally representative nature. However, the system functions only at the individual level rather than at the family level and is not publicly displayed. UKHLS has 30 weights, while PSID has only 3 weights per wave—longitudinal family, longitudinal individual, and cross-sectional, with larger supplements containing default weights. When PSID provides specific weights, the study team also describes those specific weights in video tutorials. PSID's help desk, however, can always assist users confused by the topic. NACDA provides weighting information via the user guide or codebook, and requests that data providers use a field in the deposit form to represent their weighting and collection metadata. Despite its desire to review submissions to minimize the issues inherent in sharing longitudinal study data, NACDA does not have the resources to do so and must trust that the information provided is correct.

Appendix 1. Meeting Agenda

1:00 Welcome and Introductions
Brian Harris-Kojetin, Director, CNSTAT
John Phillips, Division of Behavioral and Social Research, NIA

1:15 Facilitating Greater Use of NIA Longitudinal Studies: Data Documentation
Moderator: Mick Couper, University of Michigan

National Longitudinal Survey of Youth
Alison Aughinbaugh, Bureau of Labor Statistics

Understanding Society
Alita Nandi, University of Essex, UK

Panel Study of Income Dynamics
David S Johnson and Noura Insolera, University of Michigan

Discussion
Kathryn Lavender, National Archive of Computerized Data on Aging (NACDA)

2:45 Break

3:00 Addressing User Needs with Complex Data Systems and Multiple Data Sources
Moderator: Elizabeth Stuart, Johns Hopkins University

American Economic Association Code and Data Repository
Lars Vilhuber, Cornell University

Dataverse
Ana Trisovic, Harvard University

4:30 Wrap-up and Final Discussion and Priorities for NIA to Consider

Appendix 2. List of Participants

Presenters
Alison Aughinbaugh, U.S. Bureau of Labor Statistics
Noura Elise Insolera, Panel Study of Income Dynamics (PSID), University of Michigan
David Scott Johnson, Institute for Social Research & PSID, University of Michigan
Kathryn Lavender, National Archive of Computerized Data on Aging (NACDA), University of Michigan
James McNally, NACDA, University of Michigan
Alita Nandi, Understanding Society, University of Essex, United Kingdom
Ana Trisovic, Harvard University
Lars Vilhuber, Cornell University

Committee Members
Mick Couper, University of Michigan Survey Research Center
Robert Goerge, Chapin Hall at the University of Chicago
Elizabeth Stuart, Johns Hopkins Bloomberg School of Public Health

National Institute on Aging
John Phillips, Chief, Population and Social Processes Branch, DBSR
Frank Bandiera, Program Official, Division of Behavioral and Social Research (DBSR)
Partha Bhattacharyya, Program Director, DBSR
Rosaly Correa-de-Araujo, Senior Scientific Advisor to the Division Director, Division of Geriatrics and Clinical Gerontology
Elena Fazio, Health Scientist Administrator, DBSR
Jonathan King, Senior Scientific Advisor to the Division Director, DBSR
Jennie Larkin, Deputy Director, Division of Neuroscience (DN)
Damali Martin, Program Director, Population Studies and Genetics Branch, DN
Priscilla Novak, Program Official, DBSR
Georgeanne Patmios, Senior Scientific Administrator, DBSR
Dana Plude, Deputy Director, DBSR
Eleanor Simonsick, Epidemiologist, Intramural Research Program
Jean Tiong-Koehler, Special Assistant to the Division Director, DN

CNSTAT Staff
Brian Harris-Kojetin, Director
Melissa Chiu, Deputy Director
Connie Citro, Senior Scholar
Rebecca Krone, Program Coordinator

Rose Li & Associates, Inc. (Contractor to NIA)
Christina Tricou, Science Writer
Shanna Breil, Program Coordinator

Appendix 3. Chat Transcript

From Dana Plude to Everyone: 01:16 PM

sometimes a challenge to unmute on phone

From Partha Bhattacharyya to Everyone: 01:21 PM

what percentage of 150K (or subset there off) are used by researchers and how do you track use. Can be answer later

From Jennie Larkin to Everyone: 01:37 PM

Alison's talk and found out the answer to my question before I posted it -- that DOL supports challenges that leverage the NLS data resources:
https://www.dol.gov/newsroom/releases/dol/dol20210316

From Partha Bhattacharyya to Everyone: 02:03 PM

PSID Q: How often old research is revisted? and partnership form

From Partha Bhattacharyya to Everyone: 02:09 PM

any stats by paper of variables?

paper or variable

From Jennie Larkin to Everyone: 02:39 PM

Present and Future Data Concerns from NACDA definitely resonate.

From Kathryn Marjorie Lavender to Everyone: 02:41 PM

For all projects: How often do they update past materials? And how do you alert data users to these changes?

From Elizabeth Stuart to Everyone: 02:41 PM

We might be able to address some of the questions/comments in the discussion time of the next session too; I imagine some topics will overlap!

From John Phillips to Everyone: 02:42 PM

For all: given all you do for sharing, what complaints do you get about access?

From Kathryn Marjorie Lavender to Everyone: 02:42 PM

Another question for PSID - would you be willing to share the guidelines you use to vet results for the restricted data? We are always looking to enhance our restricted data guidance and it helps to know what others are doing.

From Alison Aughinbaugh to Everyone: 02:43 PM

Response to Partha's question about variable use: NLS can see which variables are downloaded as well and we track this information.

From David Scott Johnson to Everyone: 02:44 PM

Here is the disclosure requirements page
https://simba.isr.umich.edu/restricted/docs/ContractRestrictedData/DataDisclosureEnclave.pdf

From Kathryn Marjorie Lavender to Everyone: 02:45 PM

For all - Related publication/bibliography - how are these gathered? Do you have a team dedicated to this, user submission form?

Thank you @David!

From Alison Aughinbaugh to Everyone: 02:45 PM

We've classified variables as primary, secondary, and tertiary. NLS recommends that most users screen their searches to primary. Advanced users may want to see secondary and tertiary variables as well--include check items.

From David Scott Johnson to Everyone: 02:46 PM

The ISR library helps locate publications using google scholar and web of science, PSID staff (like Noura) check and add more

From Kathryn Marjorie Lavender to Everyone: 02:46 PM

Thank you everyone

Thanks David

From James McNally to Everyone: 02:46 PM

We don't get many complaints but we do get many interesting questions that push us to think harder about data and what is available.

From Partha Bhattacharyya to Everyone: 02:48 PM

Will code sharing help these people?

From James McNally to Everyone: 02:48 PM

Code sharing could be huge. Both for examples and to address the validation crisis.

From Alison Aughinbaugh to Everyone: 02:49 PM

NLS added a rename feature in the downloaded code to read the data. It

From Elizabeth Stuart to Everyone: 02:49 PM

I hope to come back to code sharing topics in the next session so keep these questions!

From Noura Elise Insolera to Everyone: 02:49 PM

Absolutely. There are many discussion threads and stats packages providing code. We don't distribute this on our website because we would have to vet and debug code

From James McNally to Everyone: 02:50 PM

Think of how useful the old Green Sage books were for graduate students. Classic code sharing. though old school

From Partha Bhattacharyya to Everyone: 03:09 PM

Lars- who conducts reproducibility checks?

From Partha Bhattacharyya to Everyone: 03:21 PM

MC- Lars covered it. I jumped on the question

From Partha Bhattacharyya to Everyone: 03:50 PM

before codes are made available data sometimes is tweaked - how do we deal with what I say "data cleaning" sometimes lead to driving results?

From Lars Vilhuber to Everyone: 03:51 PM

We require code for all steps starting with raw data.

From Jennie Larkin to Everyone: 03:52 PM

to the point of re-usability of code -- the current NOSI released by NIH's of Data Science Strategy might be of interest: Administrative Supplements to Support Enhancement of Software Tools for Open Science https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-073.html

From Damali Martin to Everyone: 03:56 PM

resources: People and money

From Kathryn Marjorie Lavender to Everyone: 03:56 PM

Question for later - Jupyter binder - How is data security managed with respect to use of this system?

From Partha Bhattacharyya to Everyone: 03:57 PM

how do you scale at cost?

From Elizabeth Stuart to Everyone: 04:11 PM

The Journal of Statistical Software?

From David Scott Johnson to Everyone: 04:16 PM

Does the data and code need to start from the original data set. Many PSID papers create an extract and the code works on that extract to produce the results. But sometimes you cannot use the PSID data on our site to create the data extract in the paper

From John Phillips to Everyone: 04:17 PM

RAND HRS uses version numbers...

From Lars Vilhuber to Everyone: 04:17 PM

https://social-science-data-editors.github.io/guidance/Requested_information.html and template README at https://doi.org/10.5281/zenodo.4319999

From Noura Elise Insolera to Everyone: 04:17 PM

PSID has release numbers. Each file has one.

From Alison Aughinbaugh to Everyone: 04:18 PM

Each NLS download contains a version number too.

From John Phillips to Everyone: 04:18 PM

that should help with some pathing and version issues

From Mick P Couper to Everyone: 04:18 PM

Where should the responsibility lie -- data producers (as we heard from earlier today) or data users (which seems to be the focus of the current discussion)?

From Connie Citro to Everyone: 04:22 PM

In response to one of Lars' earlier points, I'm all for the idea that producers, including statistical agencies, should provide suggested citations for datasets.

From John Phillips to Everyone: 04:23 PM

Sounds like training during the PhD and perhaps refreshers could help. The training could be build on some best practices and then made a part of formal data sharing guidance...

From Kathryn Marjorie Lavender to Everyone: 04:24 PM

Thank you

From John Phillips to Everyone: 04:24 PM

Regarding "frozen" files - a nicety is that you never need to worry about longitudinal or file building discrepancies, but then everything of interest needs to be in the block file.

From Lars Vilhuber to Everyone: 04:33 PM

https://github.com/AEADataEditor/report-aea-data-editor-2020/blob/master/AEADataEditor_Report2020.pdf

last page

@mick: it's a joint responsibility. Users have to cite, and producers should make that easy. Producers carry the responsibility of making archived/curated data available.

@david / noura: why not simply a "preset" in the Data Extractor? (or several "presets")?

Suggestion: Open up the Extractor API to R/Python/Stata

From Partha Bhattacharyya to Everyone: 04:45 PM

to add to John's questions: how do we deal studies which do not produce longitudinal file but are longitudinal - reason: due to legacy issue

From David Scott Johnson to Everyone: 04:59 PM

and many economist users don't even use the weights

From Mick P Couper to Everyone: 05:00 PM

In addition to weights, accounting for complex sample design in estimates

From David Scott Johnson to Everyone: 05:02 PM

Thanks. This was great. I even took away some best practices. I'm off to another meeting.

From Kathryn Marjorie Lavender to Everyone: 05:02 PM

Thank you so much everyone!

From John Phillips to Everyone: 05:02 PM

Thanks to the panel and participants

From Damali Martin to Everyone: 05:02 PM

thank you.

From Lars Vilhuber to Everyone: 05:02 PM

Thanks to all! Very interesting, and honored to speak to you!

From Priscilla Novak to Everyone: 05:02 PM

Thanks!

From Partha Bhattacharyya to Everyone: 05:02 PM

Thank you!

From Kathryn Marjorie Lavender to Everyone: 05:02 PM

Have a great evening!

From Allison Aughinbaugh to Everyone: 05:02 PM

Thanks.