# The Good, the Bad and the Ugly of Playing a Data Custodian

*by Chiu-Chuang (Lu) Chou* [1]

**Abstract:**
The National Survey of Families and Households (NSFH) is a prominent longitudinal study on family life in the United States. NSFH has been funded by the National Institute of Child Health and Human Development (NICHD) and National Institute on Aging (NIA). The total amount of federal grant money for NSFH is 14.5 million dollars. Three waves of surveys were conducted in 1987-1988, 1992-1994 and 2001-2003. According to the ICPSR Related Literature database, there are 1,051 publications based on the NSFH data. Researchers continue to use NSFH to study family living arrangements, marriage, cohabitation, fertility, parenting relations, kin contact and economic and psychological well-being. When the NSFH funding ended in 2006, the Center for Demography of Health and Aging (CDHA) took over providing user support for NSFH. The decision was made because it is important to continue the stewardship of this prominent study. Without the expertise of the original NSFH project team, CDHA staff needs to learn the rich content of NSFH quickly to help NSFH researchers. This paper describes the challenges and rewards we have in the last few years as we play the custodial role for the complex NSFH study. Also, our effective online analysis tool for disseminating NSFH data is explained. And finally, a future plan for repurposing NSFH data is presented.

Keywords: user support, data preservation and dissemination, National Survey of Families and Households (NSFH), Center for Demography of Health and Aging (CDHA) and Better Access to Data for Global Interdisciplinary Research (BADGIR)

**Introduction**
The Center for Demography of Health and Aging (CDHA) at the University of Wisconsin -Madison is one of fourteen P30 demography centers on aging sponsored by the National Institute on Aging (Grant Number: P30 AG17266). The National Survey of Families and Households (NSFH) is one of five major projects associated with our center. NSFH was funded by the National Institute of Child Health and Human Development (NICHD) and National Institute on Aging (NIA). The total amount of federal grant money for NSFH is 14.5 million dollars. Three waves of surveys were conducted in 1987-1988, 1992-1994 and 2001-2003. NSFH data has been widely used to study family formation, marriage, cohabitation, fertility, parenting relations, kin contact and economic and psychological well-being in the last twenty years.

When the funding for NSFH ended in 2006, CDHA assumed the responsibility to maintain the NSFH project website in the spirit of good data stewardship. Even though NSFH project ended, the research community continues to conduct secondary analysis on NSFH data. We knew it was important to continue user support for NSFH. However, none of our staff was involved in the NSFH project when it was active. Both principal investigators are retired and all the project team members left. To carry out user support for NSFH, CDHA staff dove into the NSFH documents and tried to learn NSFH as much as we can.

This paper first illustrates the rich and complex content of three waves of NSFH, then continues to describe the scope and work load of our user support service, and concludes with a plan to repurpose NSFH data in the future.

**National Survey of Families and Households (NSFH)**
American families were going through some significant changes in the 1970s and early 1980s. Divorce rates were going up, more couples chose to cohabitate, more women were joining in the labor force and the fertility rate declined considerably (Bumpass, 1990). All of these changes were affecting family life and household structure. The Federal government had a need to understand these social changes and formulate public policies to address them. In 1983, the Center for Population Research of the National Institute of Child Health and Human Development (NICHD) issued a Request for Proposal (REF No. NICHD-DBS-83-8) for a large scale data collection to study the causes and consequences of the changes happening in the families and households in the U.S. In June of 1983, a research team consisting of Larry Bumpass, James Sweet, Maurice MacDonald, Sara McLanahan, Annemette Sorensen, and Elizabeth Thomson at the University of Wisconsin-Madison sent in a proposal to design a national study covering many aspects of family experiences and related life course events. This study would have a very broad scope and a comprehensive coverage on family living arrangements,

kin contact, life histories including marriage, cohabitation, education and fertility, employment, economic and psychological well-being. The research team wanted to provide a rich data resource for comprehensive analyses of family experience from various theoretical perspectives.

In January of 1984, the Wisconsin proposal was accepted and the research team spent 18 months developing a basic design for the survey and possible question sequences. In June of 1985, Bumpass and Sweet submitted a grant proposal to the National Institute of Child Health and Human Development for the implementation of the National Survey of Families and Households (NSFH). A three-year, $4.8 million grant (R01HD021009) was awarded to undertake this cross-section survey, which was also designed to be the first round of a longitudinal study.

NSFH research team had specific plans to address the limits of other large cross-sectional surveys such as, the Panel Study of Income Dynamics (PSID), the National Longitudinal Surveys of Labor Market Experience (NLS), the Survey of Income and Program Participation (SIPP) and the Current Population Survey (CPS). NSFH sample would represent the population of the entire United States and its sample size would be large enough to permit analysis of family process and structure among subpopulations. NSFH would have an exclusive focus on family issues covering a broad range of family structure, process and relationships to allow examination of the relations among them. It would cover issues important to several disciplines and sub-disciplines and permit testing of different hypotheses related to various aspects of the American family. The survey would collect respondents' early experiences with their family and in other facets of life. Such information would be the baseline for a longitudinal study of the causes and effects of family transitions and experiences.

The main NSFH sample was a national, multi-stage area probability sample containing about 17,000 housing units drawn from 100 sampling areas in the 48 contiguous states in the U.S. The sample included a main cross-section sample of 9,643 households. The oversample of blacks, Puerto Ricans, Mexican Americans, single-parent families and families with stepchildren, cohabiting couples and recently married was accomplished by doubling the number of households selected within the 100 sampling areas. One adult (19 and older) was randomly selected from each household as the primary respondent. Face to face person interviews were conducted either in English or Spanish with portions of the interview using self-administered questionnaires. The interview lasted about one hour and forty minutes. A shorter self-administered questionnaire was given to the spouse or cohabiting partner of the primary respondent. Field work was conducted from March of 1987 to May of 1988 by the Institute for Survey Research at Temple University. There were 13,017 respondents completed wave 1 interviews.

A re-interview of NSFH wave 1 respondents was conducted in 1992-1994 with a budget of 5.2 million. The sample of NSFH2 consisted of wave 1 main respondents (N=10,007), their current spouse/partner (N=5,643), their original spouse/partner (N=789) (if the respondents parted with their spouse/partner in wave 1), their focal children (N=2,495), their parents (N=3,347) and their proxy (N=802) (when the wave1 respondent was deceased or too ill to be interviewed.) The main respondents, their current spouse/partner, and their original spouse/partner were interviewed with Computer Assisted Personal Interviewing (CAPI) technology using laptop computers. Focal children, parent and proxy interviews were done over phones with Computer Assisted Telephone Interviewing (CATI) technique.

This five-year follow-up was designed to collect information on life events including marriages, divorces, births, work experiences and other transitions since the first interview in 1987-1988. The consequences of family experiences, characteristics, attitudes, health, well-being, kinship, parenting, spousal relationship, social support, labor force participation, income sources, assets and debt enrich NSFH data for further examination on various factors influencing family dynamics, marriage, cohabitation, childbearing and marital disruption. Wave 2 data combined with wave 1 data allow researchers to study union formation and dissolution with the implications of cohabitation. Attitudes and behaviors from both male and female respondents provide a detailed picture on the transition from cohabitation to marriage, factors and effects of nonmarital childbearing among cohabitated couples, and differences on union stability.

The third wave of interviews was conducted in 2001-2003 by the University of Wisconsin-Madison Survey Center using Computer Assisted Telephone Interview technology. Due to budget constraint (4.5 million), only a subset of the original sample was interviewed. The subset consists of main respondents 45 and older by January 1, 2001 with no focal children and respondents who have a focal child, wave 1 spouses or partners of main respondents and respondents' eligible focal children (age 18-34 in wave 3). For main respondents who were deceased or too ill to be interviewed and who did not have a spouse/partner to be interviewed, proxy interviews were conducted. The instruments for respondents and their spouses/partners were the same. The focal child's questionnaire was shorter and very similar to that used in the older focal child interview in wave 2. The proxy interview used the same instrument from the wave 2 proxy interview. When all the interviews were completed, information on the 2nd follow-up survey was collected from 4,600 respondents, 2,677 spouses/partners, 1,952 focal children and 924 proxy interviews. Information collected in wave 3 included living arrangements, histories of marriage, cohabitation, education, fertility, employment, marital and parenting relationships, kin contact, and economic and psychological well-being.

### Dissemination of NSFH Data and Documentation Files

NSFH data and document files are available from several places: NSFH project website (http://www.ssc.wisc.edu/nsfh), Inter-University Consortium for Political and Social Research (ICPSR) (http://www.icprs.umich.edu/), Sociometics (http://www.socio.com), and Better Access to Data for Global Interdisciplinary Research (BADGIR) (http://nesstar.ssc.wisc.edu/webview/). Each venue has its merits and caveats.

The NSFH project website (http://www.ssc.wisc.edu/nsfh/) was created by the principal investigators to disseminate data and provide relevant information for interested users. An outline of major topics is described for each wave. So users can browse the rich NSFH content by broad categories. In addition to data, documentation files such as, sample design, methodology, codebooks, layout/dictionary files, appendices, questionnaires and skip patterns, working papers, bibliography, and FAQ for data users can be accessed at this website. Data files from waves 1 and 2 are distributed in raw ASCII format. Users need to use the layout/dictionary files to write their own setup files to extract data. The Wave 3 data files are in SPSS sav format only. Users can download them to SPSS without writing their own setup files. However, users who prefer data in other formats are responsible for data conversion. All three waves of the NSFH data and documentation files are freely available there. This site has the most comprehensive coverage of NSFH both in data and documents. The ASCII format of waves 1 and 2 data and the lack of statistical package setup files mean that users need to consult the codebook and dictionary files to create their own subsets.

NSFH Waves 1 and 2 were deposited with the Inter-University Consortium for Political and Social Research (ICPSR) in 1994 and 1997 respectively. ICPSR assigned study number 6041 to wave 1, study number 6906 to waves 1 and 2.  For wave 3, ICPSR provides only the study description (study # 171) and directs users to the NSFH project website for data access. ICPSR provides SAS and SPSS setup files for wave 1 data, but no SAS or SPSS setup files are available for the wave 2 data files. The coverage of NSFH at ICPSR is limited. However, users who are interested in publications based on the NSFH data can search the ICPSR Related Literature database at this URL: http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/index.jsp.

Sociometics has NSFH waves 1 and 2 available from their American Family Data Archive (AFDA) for a fee. Users can obtain SAS and SPSS setup files for both waves of data. The Sociometrics Multivariate Interactive Data Analysis System (MIDAS) is a subscription-based online analysis tool that provides access to NSFH data and documents. They don't have wave 3 files. The limited coverage and fee-based service might not appeal to some NSFH users.

CDHA Data Archivist, Janet Eisenhauer Smith created an online archive called Better Access to Data for Global Interdisciplinary Research (BADGIR) in 2004. BADGIR is powered by the Nesstar Suite which implements metadata developed by the Data Documentation Initiative (DDI). That same year, NSFH waves 1 and 2 data and documentation were marked up and published to the BADGIR catalog. Users can browse and search the rich content of NSFH via a friendly web interface.  Simple statistical analyses like cross-tabulation and regression can be run interactively. Custom extracts in SAS, SPSS, STATA, Excel and raw ASCII format can be done efficiently using data download feature. This online data tool gives users the flexibility to output NSFH data in a format they can use effectively. The number of wave 2 data files in BADGIR is smaller than those available from the NSFH website and ICPSR. This is because when CDHA staff marked up wave 2 data files, variables from the self-enumerated data files and constructed variables were combined in the main files for respondent, spouse and ex-spouse. In fall of 2008 we added waves 3 files to BADGIR. Because we had an access to the original CAPI questionnaires, the wording of a question pertinent to each variable is included in the metadata description in wave 3. Study methodology, sample design and other important documents are embedded in the metadata and might not be obvious to users if they access the variables directly.

### User Support for NSFH

The NSFH project included a user support component while it was funded. By August of 2006, only one graduate student was employed to provide user support. Because NSFH is such a rich data source for issues related to family, researchers continue to conduct their secondary analysis of NSFH. Hence, there is an ongoing need to provide user support. Since CDHA has NSFH in its BADGIR online archive, a decision was made by the center to continue user support when the NSFH project funding ran out. CDHA staff also assumed the responsibility of maintaining the NSFH website and continued the geomerge service for users who need to include geographical characteristics in their analysis of NSFH data.

We started to log NSFH user support questions, when CDHA merged with the Data and Program Library Services (DPLS) and the Center for Demography and Ecology (CDE) in January of 2007. Our new organization is called the Data and Information Service Center (DISC) and we used Libstats to keep track of our reference services. Libstats is an online utility developed by Eric Larson and Nathan Vack at the General Library System in the University of Wisconsin-Madison. Libstats classifies the extent of reference questions by the time our staff spent answering them. There are four categories: 0-9 minutes, 30 minutes, a couple hours and substantial. These categories are very simple and yet can reflect the complexity level of each reference question DISC staff handles. From January 1st of 2007 to June 30th of 2010, we have answered 313 questions from NSFH users. Some questions can be answered fairly quickly but some took significant time from the CDHA staff.

These three tables were created to illustrate types of patron who seek NSFH user support, the amount of time our staff spent to answer their questions and the format of their questions. The majority of questions arrived in NSFHelp email account.

| Patron Types | Researchers-Non UW-Madison | Undergraduates UW-Madison | Graduates UW-Madison | CDHA Affiliates | Communities |
|---|---|---|---|---|---|
| Counts | 285 | 2 | 9 | 3 | 14 |

**Table 1**: Questions by Patron Types (Communities include mass media and non-research related inquiries.)
Most questions came from researchers outside of UW-Madison.

| Time Spent | 0-9 minutes | 30 minutes | a couple hours | Substantial/more than two hours |
|---|---|---|---|---|
| Counts | 132 | 96 | 17 | 68 |

Table 2: Time Spent by CDHA Staff to Answer Users' Questions
42% of them were answered in less than 10 minutes

| Question Formats | Walk-up | Email | phone |
|---|---|---|---|
| Counts | 6 | 274 | 33 |

Table 3 Question Formats
The majority of questions arrived in NSFHhelp email account. We do not conduct reference services using Instant Message or Online Chat.

NSFH is a prominent longitudinal study of family life and dynamics. Its broad scope and national sample size make it a unique data source for research. Twenty years after its first wave of data collection, research papers are still being published based on NSFH data. According to the ICPSR Related Literature database (http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/index.jsp) on July 20, 2010, there are 1,051 publications based on NSFH data: 190 theses, 616 journal articles, 189 reports, 34 book sections, 11 conference proceedings and 11 books. Such high numbers of publications clearly demonstrate the strength and richness of NSFH.

The most significant contribution of CDHA is to freely disseminate public-use NSFH data and documentation online via our BADGIR catalog. CDHA staff marked up NSFH study using Nesstar Publisher with the metadata elements established by the Data Documentation Initiative (DDI) in 2004. These DDI compliant data files were published in the BADGIR catalog which a Nesstar server powers. In fall of 2008, all three waves of NSFH data have been described and documented down to the variable level in BADGIR. Table 4 shows the number of files, cases, and variables available from BADGIR. There users can search any of the 26,927 variables of their interest, run cross tabulations and regression analyses, and download customized subsets in formats like SAS, SPSS, Stata, comma delimited, and spreadsheet for further analysis. Such value-added enhancements gives users with various levels of experience an exceptionally easy point of access to the complex and rich NSFH data. (There is one caveat regarding SAS. BADGIR actually creates a SAS statement file and an ASCII data file not a SAS system file. This is because SAS did not release its proprietary codes to the Nesstar developers.)

### Restricted Geomerge Files for NSFH Users

In response to users who need to include geographical characteristics in their analyses of NSFH data, we have created geo files which can be matched with users' contextual data. To obtain geomerge data, a NSFH user first fills out a confidentiality agreement form and submits it for review by the NSFH Principal Investigator, Larry Bumpass. Currently, we only offer geomerge for waves 1 and 2. The NSFH geo file consists of the following geographic units: state, zip code, state FIPS, county FIPS, place FIPS, MCD FIPS, MSA FIPS, 1990 Census tract, 1990 Census Block, 1990 Census Labor Market and Area from MABLE/Geocorr V2.5 Geographic Correspondence Engine. Users choose the geographic unit and locate the contextual variables for that unit from other sources, such as the County and City Data Book. They then send us a copy of their contextual file in

| Wave | File Description | Number of Cases | Number of Variables |
|---|---|---|---|
| 1 | Main Respondent | 13,007 | 4,355 |
| 2 | Main Respondent | 10,005 | 4,887 |
| 2 | Spouse | 5,624 | 4,887 |
| 2 | Ex-Spouse | 789 | 4,883 |
| 2 | Proxy | 802 | 84 |
| 2 | Parent | 3,347 | 1,192 |
| 2 | Focal Child (Age 10-17 | 1,415 | 179 |
| 2 | Focal Child (Age 18-23 | 1,090 | 717 |
| 3 | Best Measures | 10,211 | 444 |
| 3 | Main Respondent and Spouse Combined | 7,277 | 2,317 |
| 3 | Roster 1 (Household Members) | 7,277 | 308 |
| 3 | Roster 2 (Sons and Daughters Living Elsewhere) | 5,224 | 512 |
| 3 | Roster 3 (Spouse/Partner's Sons/Daughters Living Elsewhere) | 1,122 | 243 |
| 3 | Marriage History | 4,418 | 133 |
| 3 | Union History | 4,418 | 177 |
| 3 | T1T2T3 Status | 13,007 | 54 |
| 3 | Focal Child Interview | 1,952 | 1,208 |
| 3 | Focal Child Household Roster | 1,670 | 125 |
| 3 | Focal Child Marriage History | 1,952 | 55 |
| 3 | Focal Child Union History | 1,952 | 97 |
| 3 | Proxy Interview | 924 | 70 |
| Total | | 97,484 | 26,927 |

**Table 4**: File Specification for NSFH data in BADGIR

raw ASCII format, a dictionary file listing the variables and their column locations, and a file with descriptive statistics for each variable in the contextual data file. My colleagues, Charlie Fiss and Janet Eisenhauser Smith review these files and merge the contextual data with individual records in the restricted geo NSFH files. At the end, a geomerge file with case ID and the characteristics of places of residence by the geographic unit chosen by the user but not geographical locations is sent to the user. This way, NSFH respondents' confidentiality is protected while researchers can add geographical aspects to their analyses of NSFH data.

**Challenges of User Support**

One of the downsides of having data and documentation accessible down to the variable level in BADGIR is that it has created high expectations from NSFH users. Some users are very anxious and jump too soon to the data without spending adequate time learning about the survey design, sampling and study methodology. This means that CDHA staff receives many NSFH questions which could be answered by the documentation available from the website. Users may miss this information when they readily access NSFH data from BADGIR.

Because none of the CDHA staff members were involved with the NSFH project when it was active, we have spent hours reading its documentation on methodology, survey design, codebook and many supplementary documents to get familiar with its rich content. It is not uncommon for staff to re-read the same document many times over when we try to answer users' questions because of the study complexities. On some occasions, we have gone to the PI for clarification and additional information. We are very careful to only provide the facts gleaned from the NSFH documentation and avoid giving advice on research methods to data users. CDHA staff does not have the expertise to discuss what variables to use in users' analyses. Staff makes this clear to users who attempt to involve staff in their research

Most of the NSFH questions we answered are from researchers like faculty, academic staff and graduate students. The experience levels and research skills of these data users vary. Graduate students who are new to secondary data analysis usually seek guidance about locating relevant sections of the codebook, questionnaires or skip patterns. More experienced users tend to ask for clarification on some coding issues for variables over time. Sometimes, a few ambitious undergraduate students have emailed us about their plans to use NSFH for a class project. When such requests arrive, staff explains to them that NSFH is probably too complicated for a class paper and suggests alternative data instead.

Because of budget constraint, wave 3 data lacks some important attributes. A set of constructed income variables were created in waves 1 and 2. Users who use wave 3 data are disappointed to learn that there is no plan to create constructed income variables for this last wave. Weight variables were not created for wave 3 files because the respondents were selected arbitrarily by their age (respondents who are older than 45 by January 21 of 2001). Without weight variables, the wave 3 files are not nationally representative. There is no solution to this problem that is being planned. These limitations have reduced the value of the wave 3 data.

A few users have reported data and documentation discrepancies to staff and we have made corrections accordingly. For example, a user reported to us that there was invalid data in his SAS data file for NSFH wave 3 respondent and spouse combined file which he downloaded from BADGIR. Our investigation revealed that those invalid values are related to income variables. When household income is over one million dollars, the export engine in BADGIR does not allocate enough columns when the requested output format is SAS. We increased the column width for income variables and republished the data file in BADGIR.

Occasionally, CDHA staff receives questions from readers of mass media because NSFH is mentioned in a recent article in a popular magazine or news story. On 2008 Father's Day, Lisa Belkin in her New York Times magazine article, "When Mom and Dad Share It All", mentioned that, "The most recent figures from the University of Wisconsin's National Survey of Families and Households show that the average wife does 31 hours of housework a week while the average husband does 14 — a ratio of slightly more than two to one." We got many calls and emails from readers of Ms. Belkin's article. They asked us to verify the hours spent on household works by husband and wife reported in that article. Those numbers were likely produced by researches using NSFH data -- but because we don't know who did the secondary data analysis or what variables were used, staff cannot recreate the analysis. Thus, we suggested those inquirers to contact Ms. Belkin directly for assistance.

## Future Plans for NSFH Data

The long-term plan is to create longitudinal NSFH files combining all three waves of NSFH data. These files will be very useful for research on the changes in family and household and the associated causes and effects over time. A match of NSFH data with the National Death Index is also being considered. CDHA will archive the address files for wave 3 and create a geocode file for these addresses. All these plans will be carried out if CDHA has sufficient funding in our third round of grant from the National Institute of Aging (Grant Number: P30 AG17266: 2009-2014).

## Conclusion

Research based on the NSFH surveys has contributed significantly to the understanding of causes and consequences of changes in U.S. families and households in the last 20 years. Multiple access points to NSFH such as, ICPSR, Sociometics, the NSFH website, and BADGIR give users different options for obtaining its data and documents. BADGIR, an enhanced discovery tool, has made access to NSFH easier. It is clearly an efficient web-based dissemination model for the NSFH study. Its free, versatile and friendly user interface has drawn attentions from professors who use NSFH in their classes. One of them is Rachel Gordon, an Associate Professor in the Department of Sociology and Institute of Government and Public Affairs in the University of Illinois at Chicago. In her text book titled Regression Analysis for the Social Sciences published in 2010 by Routledge, she includes examples from NSFH data analyses using BADGIR. We are pleased that NSFH and BADGIR utility are being used in this new textbook to advance the learning in secondary data research.

It has been four years since our center became the custodian of the NSFH study. We will continue to provide user support to the research community in the spirit of good stewardship of a milestone social science study.
It is rewarding to know that our assistance to NSFH researchers is appreciated. Rarely our good intention is damped by aggressive users who want us to do their work for them. Keeping good communication with NSFH users is important to us. We know our continuing support to the NSFH research community is relevant and important. With BADGIR we are confident to deliver reliable and informative user support for NSFH community.

## References

Bumpass, Larry L. 1990. "What's Happening to the Family? Interactions Between Demographic and Institutional Change" Demography, 27(4): 483-498.

Sweet, James, Larry Bumpass, and Vaughn Call. 1988. The Design and Content of the National Survey of Families and Households. http://www.ssc.wisc.edu/cde/nsfhwp/nsfh1.pdf.

Trull, Elaine, Lisa Famularo. 1996 National Survey of Families and Households Wave 2 Field Report. ftp://elaine.ssc.wisc.edu/pub/nsfh/cmapp_n.001.

Wright, Debra. 2003. National Survey of Families and Households Wave 3 Field Report. http://www.ssc.wisc.edu/nsfh/wave3/fieldreport.doc.

Introduction to the NSFH1 Codebooks and Other Documentation. http://www.ssc.wisc.edu/nsfh/intro.htm.

Introduction to the NSFH2 Codebooks and Other Documentation. http://www.ssc.wisc.edu/nsfh/intro.htm.

Introduction to the NSFH3 Codebooks and Other Documentation. http://www.ssc.wisc.edu/nsfh/intro.htm.

ICPSR Related Literature Database. http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/index.jsp.

**Note**

[1] Chiu-Chuang (Lu) Chou, Senior Special Librarian, Data and Information Services Center and Center for Demography of Health and Aging, University of Wisconsin Madison, email: cchou2@wisc.edu, 3308 Social Science Building, 1180 Observatory Drive, Madison, WI 53706, USA.