## ANECDOTES

# Challenges of Large-Scale Data Processing in the 1990s: The IPUMS Experience

Diana L. Magnuson and Steven Ruggles, *Institute for Social Research and Data Innovation, University of Minnesota, Minneapolis, MN, 55455, USA*

*When it was launched in 1991, the Integrated Public Use Microdata Series (IPUMS) project faced a challenging environment and limited resources. Few datasets were interoperable and much data collected at great public expense was inaccessible to most researchers. Documentation of datasets was nonstandardized, incomplete, and inadequate for automated processing. With insufficient attention to preservation, valuable scientific data were disappearing (see Bogue et al., 1976). IPUMS was established to address these critical issues. At the outset, IPUMS faced daunting barriers of inadequate data processing, storage, and network capacity. This anecdote describes the improvised computational infrastructure developed in the decade from 1989 to 1999 to process, manage, and disseminate the world's largest population datasets. We use a combination of archival sources, interviews, and our own memories to trace the development of the IPUMS computing environment during a period of explosive technical innovation. The development of IPUMS is part of a larger story of the development of social science infrastructure in the late 20th century and its contribution to democratizing data access.*

The U.S. Census Bureau played a key role in the development of computing technology. The two leading computer companies of the middle decades of the 20th century—IBM and Remington Rand's UNIVAC division—had roots as data processing companies that built equipment for the Census Bureau, and the Bureau indirectly funded the development of the first commercial computer beginning in 1946 [26].

The 1960 census was fully computerized, and the Census Bureau began to publish data in machine-readable form. In 1962, the Census Bureau introduced the first machine-readable microdata file, a one-in-1000 sample of households drawn from the 1960 census. Following the 1970 census, the Census Bureau greatly expanded the quantity of microdata, providing six 1% samples from the 1970 census—a 60-fold increase compared with 1960.

After the 1970 samples were released, the Census Bureau contracted to produce a new version of the 1960 microdata, expanded from 1-in-1000 sample density to 1-in-100. The 1960 work was carried out by the Data Use and Access Laboratories (DUALabs), another Census Bureau spin-off, which made the decision to adopt the coding schemes and record layout that had been used for the 1970 samples [24].

The availability of two compatible census years led to an explosion of research in the 1970s on social and economic changes over the course of the 1960s. The pair of microdata samples became essential tools of American social scientists. It was in this climate that two demographers proposed extending the data series backward by digitizing census microdata from earlier years. The enumeration manuscripts for the 1900 census were made accessible to scholars by the National Archives and Records Administration in 1972. Between 1976 and 1980, demographer Samuel Preston of the University of Washington directed a project funded by the National Science Foundation (NSF) to

create a one-in-750 sample of the 1900 census. Halliman Winsborough, a demographer at the University of Wisconsin, independently came up with the idea for creating samples of the 1940 and 1950 censuses. The 1940 and 1950 censuses were still protected by census confidentiality rules, so the samples were created within the Census Bureau between 1978 and 1984 with funding from an NSF grant. Preston moved to the University of Pennsylvania in 1979 and, with the release of the 1910 census manuscripts in 1982, he led a project between 1983 and 1989 to create a sample of the 1910 census with funding from both NSF and the National Institute of Child Health and Human Development (NICHD) [8], [10], [30], [34], [38], [39].

At the University of Minnesota, the use of computer technology to understand historical transformations through quantitative methods began in the decade between 1976 and 1985, the same period that the first historical microdata samples were being created. During this era, academic computing at the University was centralized on a mainframe computer in the nearby suburb of Lauderdale, Minnesota. Although punch cards remained the primary means of creating datasets, interactive computing through teletype machines became available in the mid-1970s [17], [36].

Historical demographers Robert McCaa (hired in 1974), Russell R. Menard (hired in 1976), and Steven Ruggles (hired in 1984) undertook collaborative work that contributed to the expansion and redefinition of social science data infrastructure. All three had been working with historical census microdata since the mid-1970s, using card punches to digitize historical enumeration manuscripts from Ramsey County, Minnesota (Menard) [4], Parral, Mexico (McCaa) [15], and Lancashire, England (Ruggles) [19].

By the mid-1980s, the limitations of the mainframe computing system for large-scale historical datasets were apparent. Shortly after he arrived at Minnesota, Ruggles began working on the Control Data Corporation CYBER mainframe computer using the newly created 1940 Public Use Microdata Sample. He used the Network Operating System "COPY" command to transfer the data files from a temporary disk to a permanent one. The charge for executing that single command was $500, using up half of his annual allocation of computer time.[1]

To avoid the high cost of the centralized computer system, the historians turned to microcomputers. McCaa had been a microcomputer enthusiast since they were introduced in the 1970s and by the early 1980s was using microcomputers for census data entry in Mexico [15], [16].[2] In 1985, Menard and Ruggles established the History Micro-Computer Lab—later known as the Social History Research Laboratory (SHRL)—with a grant from IBM. The lab initially had two IBM computers, but soon expanded to five, including the latest model, the IBM PC/AT. SHRL obtained and made available a wide range of historical datasets for teaching and research.

> During the 1985–1986 academic year, we made available for classroom use: a large national sample from the 1900 federal population census; data from the population and agricultural census schedules for rural Ramsey County, Minnesota, for 1860 through 1880; biographical information on 1500 graduates of Yale College in the eighteenth century; and information on basic demographic parameters of early America that permits exploration of how migration and vital rates interacted to shape the growth of population. During 1986-87, we developed additional data sets, particularly census material from northern Mexico during the late eighteenth and early nineteenth centuries; European family reconstitution data from 1600 to 1800; and evidence on the family budgets of industrial workers in Germany and the United States during the 1890s; national samples of the federal census for the years 1910, 1960, 1970, and 1980; and census returns for the Red Lake, Minnesota Indian Reservation for 1900 and 1910 [35, pp. 70–71].

The 1900, 1910, 1960, 1970, and 1980 census samples made available in the SHRL were not the full samples, since the full datasets were far too large to be used on a microcomputer. Rather, they were tiny subsamples, with 2100 households (6000 to 10,000 persons) from each census. The earliest complete microdata sample was transferred to a microcomputer in 1987, when graduate student Ron Goeken was assigned the task of breaking the 1900 public use sample into pieces small enough to fit on floppy disks. Goeken then copied the files to Ruggles' microcomputer and reassembled them. The full sample was just under 10 MB, consuming half the computer's available disk space.[3]

---

[1]Steven Ruggles, interviewed by D. L. Magnuson, University of Minnesota, Jan. 9, 2014.

[2]Robert McCaa, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Dec. 6, 2021.
[3]Ronald Goeken, interviewed by D. L. Magnuson, University of Minnesota, Oct. 3, 2013; Ronald Goeken, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 6, 2021.

Ruggles had close connections to the Preston and Winsborough census projects. He was an undergraduate at the University of Wisconsin in 1978 when the 1940–1950 project was launched, and he was a postdoc there in 1984 when it was completed. In between, he was a graduate student at the University of Pennsylvania during the creation of the 1910 project, and he was among the earliest users of the 1900 microdata sample, which formed the basis for much of his dissertation research.

By 1989, Preston and Winsborough had both decided that they were done with census digitization projects, and they were happy to share their grant proposals. Borrowing the strongest parts of those, Ruggles and Menard sought NICHD funding to create a 1-in-500 public use microdata sample of the 1880 U.S. census [12], [27], [29].

The 1880 census was the first to collect data on marital status, family relationships, and parental birthplaces, making it especially useful for studies of household composition, fertility, and nuptiality [27]. Ruggles' research focused on long-run changes in family and household composition, including multigenerational households, boarding and lodging, single parenthood, and living alone. Even though the Census Office began collecting full information on family and household composition in 1880, it did not publish any statistics on the topic until 1940, and only a few rudimentary measures thereafter. The series of microdata files, therefore, offered unprecedented opportunities to examine changes in families and households for the country as a whole. Since the 1880 Census was the first to identify family relationships and marital status, it made a natural starting point for studies of long-run family change.

The 1880 proposal earned a priority score in the top 1%. The reviewers had one major reservation: They felt that the sample size was too small. Accordingly, the panel recommended funding only if the proposal were revised from the proposed 1-in-500 density to 1-in-100, with an appropriate adjustment of the budget. Ruggles and Menard enthusiastically agreed to make the change.

According to Ruggles, he and Menard did not know if "the first one [1880] would fly," but once it did, they were confident that they could secure funding and "fill in the rest."[4] Their confidence was not misplaced: Federal funding for the creation and dissemination of new samples for the 1850, 1860, 1870, 1900, 1910, 1920, and 1930 censuses came to the University of Minnesota between 1992 and 2002.

## TECHNOLOGICAL ENVIRONMENT, 1989–1994

The University of Minnesota history department shared access to the Social Science Research Facilities Center (SSRFC) with other social science departments at the university. Located on the West Bank in the basement of Blegen Hall, SSRFC was organized in the 1964–1966 biennium and designed to facilitate and support research at the university [37]. SSRFC was designed to provide access to the mainframe computer in Lauderdale (approximately 5 mi/8 km from Blegen Hall), and housed keypunch machines, a card reader, a line printer, and a punch card sorting machine. As interactive computing became feasible in the later 1970s, SSRFC installed computer terminals; in the 1980s, SSRFC added a PDP-10 minicomputer and nine-track tape drive.[5]

SSRFC staff provided essential technical computing support for the 1880 census project (and subsequent census projects) and the Integrated Public Use Microdata Series (IPUMS) during the 1990s.[6] To complement the microcomputers in the History Department's SHRL, Ruggles and Menard began to assemble the computational infrastructure necessary to support the 1880 PUMS project.

The 1880 Public Use Microdata Sample grant (1989–1994) provided funding to purchase motorized Dukane 35-mm microfilm readers, data entry equipment, data processing equipment, and software. Computer Marketing Corporation 80386 microcomputers for data entry were selected through a competitive bidding process. Workstations were connected by a 2-Mb Ethernet to a Sun SPARCstation file server and a 9-track tape unit. Microcomputers configured as data entry terminals were directly networked with the Sun, which eliminated file transfers via modem or floppy disk. Graduate research assistant Bill Block owned a drill and came in one evening to drill holes through the concrete-block walls of four adjacent offices to connect the data entry terminals with the SPARCstation [11].[7]

Office space was provided by the History Department at the University of Minnesota, where the project was initially located. The 1880 project was

[4]Ruggles interview, Jan. 9, 2014.

[5]Phil Voxland, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 21, 2021. John Easton, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 21, 2021.
[6]Phil Voxland, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 21, 2021. John Easton, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 21, 2021.
[7]William Block, interviewed by D. L. Magnuson, University of Minnesota, Dec. 19, 2013.

**FIGURE 1.** Graduate Research Assistant Office, 1993.

allocated a corner office to start. Graduate research assistants also made use of Ruggles' office, conveniently adjacent to the corner office. As the project work expanded in the early 1990s to include 1850 and 1920 PUMS and IPUMS, a second office was allocated by the department at the opposite corner of the same hallway; this space was used exclusively by graduate research assistants. The space was extremely tight (see Figure 1). The first SPARCstation server was named "legohead" because space was so crowded that graduate students should have had Legos on their heads so they could sit on top of one another [13], [14]. There was no funding for office furniture, so project personnel obtained old furnishings from the university's warehouse, with most items dating back to the 1940s or 1950s. Graduate research assistant Dan Kallgren remembered,

> "We ended up getting stone-lined fireproof cabinets that were designed to hold index cards, but they were perfect for holding boxes of microfilm reels. The cabinets weighed hundreds of pounds apiece. We somehow manhandled them onto the back of my pickup truck and brought them over to the Social Science Tower... and we ended up with this weird collection of desks and stuff for the research assistants' office."[8]

In the 1990s era of computing technology, one could not separate the room from the technology.[9] The data entry operators (DEOs) were civil servants, and each had a designated desk with a desktop computer and microfilm reader. Because office space was tight, DEOs shared their workspace with graduate research assistants associated with the historical census projects. The workstations for graduate research assistants (RAs) were undesignated spaces that were used at all hours of the day and night for their work to verify the data and research procedural histories. It was understood that anyone could use any workstation, but in practice, grad RAs tended to make use of a favorite spot based on the time of day they were in the office. Shared office space created a productive synergy between and among DEOs and graduate research assistants (shown in Figure 2).

With a couple of exceptions, most graduate students arriving in the history department had limited computer experience. Lisa Dillon, Diana Magnuson, and Dave Ryden remember a supportive office culture of shared learning around computing skills and data construction. Graduate research assistants learned how to use the software from each other. Lisa Dillon recalled,

> "I learned what I learned by sitting next to people and people showing me... it was [Matt]

---

[8]Daniel Kallgren, interviewed by D. L. Magnuson, University of Minnesota, Oct. 31, 2013.

[9]Lisa Dillon, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Feb. 4, 2022.

**FIGURE 2.** IPUMS Staff, 1996.

Sobek who showed me what a spreadsheet was and how it worked."

Todd Gardner trained Diana Magnuson in the use of Norton Commander, a file-managing tool. Being "in the room where it happens" was critical to quickly skilling up and becoming a productive and creative member of the historical census projects and IPUMS. As evidence of this early era of computing technology, several members distinctly remember the very first electronic mail they saw and sent; an email exchange with Steve Ruggles, checking in on the grad RAs while he was on sabbatical in England in the spring of 1990.[10]

In addition to their data production tasks, graduate research assistants carried much of the administrative load in the early years of the 1880 project through the launch of IPUMS in the early 1990s. The history department was not accustomed to grant administration and did not have adequate capacity to meet the growing administrative needs of the projects. History department personnel were available for guidance and troubleshooting, but day-to-day administrative details were carried out by graduate research assistants in consultation with Ruggles and Menard. At the startup of the 1880 PUMS project, Dan Kallgren managed administrative duties that included coordinating purchasing most hardware and software assets for the project.

Additionally, he researched used commercial coffee makers, made his purchase from a restaurant supply company in St. Paul, and transported the large coffee maker, strapped to the back of his bicycle, back to the Social Science Tower on the West Bank. Coffee fueled DEOs, graduate students, and faculty alike, and the machine served as the locus of a community building and gathering space. David Ryden administered tasks related to the 1920 PUMS project, which began in 1993. Ryden also had a bicycle, and this came in handy for the many trips he made back and forth across campus to obtain signatures and drop off forms related to historical census project contracts. Beginning in the fall of 1995, Cathy Fitch gradually assumed a variety of IPUMS project administrative responsibilities. The procedural history was largely researched by graduate research assistants Diana Magnuson and Miriam King, a post-doc at the university's Hubert H. Humphrey Center for Public Affairs. Thus, the computational and procedural infrastructure that developed in the 1990s was administered by graduate research assistants taking on a variety of responsibilities necessary to processing, managing, and disseminating what would become the world's largest population data collection.[11]

Software development for the 1880 PUMS was the top priority as the project got underway. Microcomputers were used for data entry, post-entry data consistency checking, and verification. The data entry software was the Integrated System for Survey

---

[10]Kallgren interview, Oct. 31, 2013; David Ryden, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Jan. 24, 2022; Catherine Fitch, interviewed by D. L. Magnuson, University of Minnesota, Jun. 16, 2015; Dillon interview, Feb. 4, 2022.

[11]Kallgren interview, Oct. 31, 2013; Ryden interview, Jan. 24, 2022; Fitch interview, Jun. 16, 2015.

Analysis (ISSA), which had been developed by West-inghouse for processing the Demographic and Health Surveys [6]. Unlike competing data-entry software, all the individual-level information for a dwelling could be included on the screen simultaneously. This meant that the DEOs could see the individual-level information for the entire dwelling while entering the data for each individual. The software had extensive capabilities for real-time consistency checks between fields; for example, it could be set up to require that married, divorced, or widowed persons were of marriageable age.

The ISSA software had significant drawbacks. The system required extensive programming. It was awkward to navigate between data entry fields except in a fixed sequence. The 1880 project sought to capture long alphanumeric strings verbatim to ensure minimal information loss, and ISSA did not handle long strings well. There were few string handling functions, and forms with many long strings were unstable and frequently crashed [3], [6].

As work progressed, it became apparent that microcomputer memory capacity was insufficient to carry out the final recoding of string variables into numeric codes. Post-entry data processing was thus shifted to the Sun UNIX workstation, which was faster and had more memory and storage capacity [6]. This operation relied on FORTRAN-77 programs and data dictionaries containing all the alphabetic entries transcribed by DEOs and their corresponding numeric codes as classified by graduate research assistants [6].

As the 1880 project was underway, Ruggles and Menard obtained additional funding to launch projects to create samples of the 1850 census in 1992 and the newly released 1920 census in 1993 [23], [28]. The 1850 project continued to use the ISSA data-entry software, but ISSA could not accommodate the expanded length of the 1920 records. After an extensive search for appropriate commercial software, the 1920 investigators decided to develop customized data-entry software using Microsoft C and Liant Software's C-scape. This approach allowed design of screen forms that mirrored the manuscript enumeration form. In addition to developing an intuitive screen layout, interactive data consistency checking for string and numeric data was built into the program [6].

## FIRST IPUMS PROJECT 1992–1995

By 1991, 10 machine-readable public use microdata samples covering the decennial censuses from 1880 and 1990 (1850, 1880, 1900, 1910, 1940, 1950, 1960, 1970, 1980, and 1990) were publicly available or under development. Use of the PUMS as a time series was impeded by incompatibilities: They were created at different times by different investigative teams, resulting in different formats, different documentation, different record layouts, and different coding schemes. The only exception was the 1960 and 1970 samples, because the new version of the 1960 sample created in 1973 had been redesigned to share the same coding and layout as the 1970 samples.

To enable an analysis of long-run family change, between 1985 and 1989, Ruggles developed a set of FORTRAN programs that recoded selected variables into a common format across the available census samples, created subsets of the samples that were of manageable size, and pooled multiple censuses into a single file [33].[12] This first effort at making variables compatible across data sets had liabilities. Initially, the guiding principle when combining the samples was a "lowest common denominator" approach [24, p. 1405]. Only a small subset of variables was converted into the common codes. The result of this method was a significant loss of information, which meant that customization of the programs was usually necessary to meet the requirements of any specific research project [24].

Despite the limitations, there was considerable demand for custom-designed, consistent-format census microdata to meet the research and teaching needs of the university's graduate students and faculty, as well as a few researchers from other institutions. By 1991, SHRL had a server dedicated to common format extracts that ran nearly continuously to meet researcher demand. The source data were stored on a set of Write-Once Read Many disks read by an IBM 3363 optical drive. Extracts for a subpopulation across all eight existing census samples required up to a day to process, and the SHRL was preparing about 25 such extracts per month [24].[13]

In 1991, Ruggles proposed to the NSF to create a single integrated series that would "maximize comparability and minimize information loss" across the eight public use microdata samples [20], [21], [22], [33, p. 103]. Ruggles also anticipated adding data for the remaining publicly available census years (1860, 1870, and 1930), thus "constitut[ing] a resource of unprecedented power for the study of long-term social

---

[12]Todd Gardner, interviewed by D. L. Magnuson, University of Minnesota, Oct. 24, 2013; Matthew Sobek, interviewed by D. L. Magnuson, University of Minnesota, Oct. 10, 2013.
[13]Sobek interview, Oct. 10, 2013.

change" [22]. Ruggles remembers the moment he went into the History Department lounge on the sixth floor of the Social Science Tower and said, "IPUMS! Integrated Public Use Microdata Series! Isn't that a great idea?" The response from the graduate research assistants was not enthusiastic.

"What a terrible name! You can't call it that!" According to Ruggles, "It was universal; everyone thought it was just a horrible name... It wasn't a bad idea to propose, just a terrible thing to call it."[14]

The 1991 NSF IPUMS grant proposal outlined four steps for

"convert[ing] the public use sample series into a single coherent form: 1) planning and design of record layouts, coding schemes, and constructed variables that maximize comparability and minimize information loss; 2) software development for reformatting, recording, constructing new variables, consistency checking, and allocation of missing and inconsistent data; 3) data processing of approximately 65 million records; and 4) preparation of an integrated set of documentation for the entire series of datasets, including a general user's guide, a volume of procedural histories, and a volume on technical characteristics and error estimation" [22].

IPUMS succeeded because of two key technical innovations of the early 1990s: 1) In 1991, IPUMS introduced the first structured metadata system for data integration; and 2) in 1995, IPUMS produced the first interactive web-based system for creating customized pooled datasets.

The IPUMS software was designed to be driven entirely by machine-processable metadata. The project adopted what is now known as a data warehousing approach, which transforms data from heterogeneous sources into a single schema. The key metadata element underlying the process was called a "translation table." A separate translation table was prepared for each variable in the system; an excerpt of the translation table for the "Relationship" variable appears in Figure 3. The top of the translation table provides the input column locations for variables for each source dataset. For example, in the original 1880 dataset, the relationship variable was on the person record and ran

---

[14]Ruggles interview, Jan. 9, 2014.



**Key**

```
RELATE.TRN
Relationship                      ← Input column range

##                                ← Data quality flag location
1880 P   21  23  79
1900 P   09  11  28
1910 P   14  16                   ← IPUMS labels
1920 P   11  14
1940 P   11  14  97               ← Standardized IPUMS codes
1950 P   16  20  63                 (general and detailed)
1960 P   01  02
1970 P   01  02 105               ← Codes in original samples
1980 P   02  04 140
1990 P   09  10 184
##

#                         IPUMS 1880 1900 1910 1940      1950 1960 1970 1980 1990
HEAD & RELATIVES (1-10):
Head/Householder          01 01  100  100  100  019901999       0-   0-   000  00
                          01 01                                 00   00
Spouse                    02 01  120  120  120  029902999       1-   1-   010  01
  Husband, not Head       02 01  140  140
  2nd/3rd Wife (PG)       02 02  121  129
Child                     03 01  130  130  130  039903999       2-   2-   020  02
                          03 01                                 20   20
  (1970 screw-ups)        03 01                                      22
  (1970 screw-ups)        03 01                                      26
  Adopted Child           03 02  132  132  132
  Stepchild               03 03  131  131  131  049904999                      03
  Adopted, ns             03 04            280
Child-in-law              04 01  133  133  133  059905999       30   30   051   *
  Step Child-in-law       04 02  134  134
Parent                    05 01  210  210  210  079907999       32   32   040  05
  Stepparent              05 02  211  211  211
```

Relate table continues for 218 additional rows ↓

Category requiring custom programming

**FIGURE 3.** Translation table.

from column 21 to column 23. The additional input field refers to the location of data quality flags in the original data, which provide information about missing or inconsistent responses.

The rest of the table has a row for each relationship code. On the left are the IPUMS labels for each category and the standardized IPUMS code. The IPUMS code comes in two parts: The first two digits are a lowest common denominator designed to be consistent across all samples, and the second two digits provide additional detail available in a subset of censuses. This composite coding system provides interoperability without losing information. The input codes are on the right of the table, with separate codes for each census year. For example, the Head/Householder category was coded 100 in the 1880–1910 censuses, 0199 in 1940, and 000 in 1980. In addition to the translation tables, other machine-processable metadata components specified the order in which variables were to be processed, the names and locations of input and output data files, and the record layout of output files [9].

Metadata-driven data integration was a major innovation. Prior data harmonization efforts used statistical packages or software code to recode multiple datasets into a common coding scheme. Such systems were cumbersome to program and difficult to maintain. The translation table provided a visual representation of recodes across all datasets simultaneously that greatly simplified the design of integrated codes. The system made it simple to develop tests of internal consistency

of the recodes (e.g., each input code must correspond to just one output code) and to ensure that all values encountered in the data appear in the translation tables. Without structured machine processable meta-data, it would have been far more expensive to build the initial IPUMS or to expand it to cover additional data sources.

The second major IPUMS innovation was the interactive web-based data extract system. The extract system was made possible by the timing of the IPUMS project coinciding with the rapidly changing computing environment of the early 1990s. When IPUMS was proposed, few social scientists had access to the Internet, and the World Wide Web existed only on a single computer in Switzerland [5]. Data transfers in that period were accomplished by sending physical media through the mail, usually 10.5-in (26.67 cm) 9-track reel-to-reel tapes.

The original IPUMS proposal called for the purchase of 450 of these tapes. A third of the tapes were needed to hold one copy of the full set of census microdata in its original format. Another 150 tapes were needed to hold the transformed data in IPUMS format. Finally, the last 150 tapes were needed to send one copy of the IPUMS-format data to the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan, which would handle the dissemination by sending copies on 9-track tapes to institutions around the country [22].

In the end, the IPUMS project purchased no tapes. In 1992 and 1993, data archivist Ann Gray at the Cornell Institute for Social and Economic Research made the Census Bureau microdata samples available for download through an anonymous file transfer protocol (FTP) server, a program that made it possible for outsiders to log on and download data. The IPUMS team was able to directly download the files over the Internet without using any tapes.

When it came time to disseminate the IPUMS data, the same technology was used. Instead of sending the data to ICPSR on 9-track tables, IPUMS set up an anonymous FTP site to disseminate the data [1], [33]; the first IPUMS dataset was downloaded on November 19, 1993.

On November 11, 1993, eight days before the first IPUMS Internet download, an undergraduate from the University of Illinois released the first successful web browser for a PC: NCSA Mosaic 1.0, thus opening broad access to the World Wide Web [18]. The first IPUMS website (see Figure 4) appeared 16 months later, one of the first 15,000 websites created. The initial IPUMS website offered the same functionality as the IPUMS FTP site: Users were able to

### Welcome to the Social History Research Lab!

**University of Minnesota**

This web site currently contains the data and documentation for the **Integrated Public Use Microdata Series (IPUMS)**. The **IPUMS** is a database consisting of 23 samples of the U.S. Census from 1850 to 1990. The **IPUMS** assigns the different samples consistent codes and integrates their documentation. At present, you may only download compressed **IPUMS** data files. To obtain a DOS decompression program click here.

If you need uncompressed data or have any other questions about the IPUMS database, contact us at ipums@atlas.socsci.umn.edu

If you have problems. questions, or suggestions about this page, send e-mail to block@torgo.hist.umn.edu

The User's Guide for the IPUMS is available online in Word 6.0 format (see below). This document is 800 pages long, and is contained in 61 separate files. A printed version is available from us for $30; contact ipums@atlas.socsci.umn.edu

**IPUMS DATA FILES:**

To download an IPUMS sample, just click on it:

1850 sample (6.9M)

1880 sample (21.1M)

1900 sample (3.6M)

1910 sample (14.0M)

1920 sample (14.4M)

1940 sample (62.4M)

1950 sample (72.8M)

1960 sample (77.4M)

1970 5% state sample (98.5M)

1970 15% state sample (98.5M)

1980 B sample (142.0M)

1990 1% sample (163.2M)

**IPUMS USER'S GUIDE:**

The following files taken together comprise the User's Guide. All files are in Word 6.0 format. As mentioned above, we will provide upon request a printed, double-sided version of the Guide for $30. If you wish to print your own copy of the Guide, all of the following files should be downloaded (just click on each in succession). Each filename begins with "guide" followed by a letter, a numeral, and then a ".doc" extension. The letter and numeral describe the order of the file within the documentation (A7, A8, B1, B2, C1, etc). For a file describing the documentation files click here.

**FIGURE 4.** First IPUMS website, March 1995.

download data and documentation files by clicking on them.

The largest file available in the 1993 version of IPUMS—the 1990 1% sample—was 163 MB. Files of that size were difficult to reliably transfer over the Internet and most social scientists found that scale of data challenging or impossible to store or process. To analyze changes over the entire period from 1850 to 1990—which was after all the point of the IPUMS—one would have to download the entire data series, 775 MB, and subset and merge the files to analyze change over time. Only investigators at major universities had the infrastructure to handle data on this scale [31], [32].

As noted, the SHRL had been creating customized harmonized data extracts for five years prior to the development of IPUMS. These extracts used a FORTRAN program to select specific variables and population subgroups needed to conduct a particular analysis and read the data from multiple files to produce a pooled extract with data from multiple censuses. Few users had the storage capacity or processing power needed to handle the very large public use microdata samples in their original form, so the process of

## IPUMS Data Extract

# Sample Selection

Reset Values | Continue To Variable Selection

What is your email address? [_____]

| Census Year | |
|---|---|
| | ☐ 1850 Sample |
| | ☐ 1880 Sample |
| | ☐ 1900 Sample |
| NOTE: IPUMS-95 contains only the 1970 5% and 15% state and county group samples, the 1980 B sample, and the 1990 1% sample. | ☐ 1910 Sample |
| | ☐ 1920 Sample |
| | ☐ 1940 Sample |
| | ☐ 1950 Sample |
| | ☐ 1960 Sample |
| | ☐ 1970 5% State Sample |
| | ☐ 1970 15% State Sample |
| | ☐ 1980 B Sample |
| | ☐ 1990 1% Sample |

| Sample Density | ● Tiny<br>○ Small<br>○ Regular |
|---|---|
| File Type | ○ Flat<br>● Hierarchical |
| Data Quality Flags | ☐ Include all data quality flags |

(ip´-ums) IPUMS

**FIGURE 5.** IPUMS data access system, 1995. Step 1.

## IPUMS Data Extract Variable Selection

Clear | Submit

**Household Variables**

| ☐ Technical Variables | | |
|---|---|---|
| ☐ RECTYP | Record type | |
| ☐ YEAR | Census year | |
| ☐ DATANUM | Data set number | |
| ☐ SERIAL | Household serial number | |
| ☐ NUMPREC | Number of person records following | |
| ☐ SUBSAMP | Subsample number | |
| ☐ HHWT | Household weight | |
| ☐ NUMPERHH | Number of person in household | |
| ☐ DWSIZE | Dwelling size | |
| ☐ NMEMBERS | Number of members in sample unit | |
| ☐ NUMHH | Number of households in dwelling | |
| ☐ NUMHHTAK | Number of households sampled from dwelling | |
| ☐ UNREL | Unrelated persons in household | |
| ☐ SLPERNUM | Sample line person number | |
| ☐ SELFWTHH | Self-weighting sample identifier | |
| ☐ Location Characteristics | | |
| ☐ REGION | Census region and division | ☐ range |
| ☐ STATEICP | State (ICPSR code) | ☐ range |
| ☐ STATEFIP | State (FIPS code) | ☐ range |
| ☐ COUNTY | County | |
| ☐ CIVDIV | Civil Division | |
| ☐ COUNTY90 | County, 1900 PUMS classification | |
| ☐ CNTYGP97 | County group, 1970 | |
| ☐ CNTYGP98 | County group, 1980 | |
| ☐ PUMA | Public use microdata area | |

**FIGURE 6.** IPUMS data access system, 1995. Step 2.

selecting variables, subsetting, and merging was essential to make the data usable.

The unlikely catalyst that broke this bottleneck was fantasy football. Graduate research assistant Todd Gardner, whose responsibilities were primarily coding and software development, was also commissioner of the graduate research assistants' office fantasy football league. Gardner was learning to use the Perl language to develop a Web application that produced box scores and fantasy football standings in a web output.[15] Unlike virtually all websites of the era, Gardner's fantasy football system was truly interactive: Users made a series of selections on a web form and the software then constructed a new customized web page based on those selections. When Gardner showed Ruggles his fantasy football website, they recognized the potential to adapt the system to create a web-based interactive front end that could operate the FORTRAN data extract program, allowing outside users to design their own customized datasets.

## IPUMS Data Extract Case Selection

Clear | Submit

**Household Variables**

| Region | ☐ New England<br>☐ Middle Atlantic<br>☐ East North Central<br>☐ West North Central<br>☐ South Atlantic<br>☐ East South Central<br>☐ West South Central<br>☐ Mountain<br>☐ Pacific<br>☐ Military/Military reservations<br>☐ PUMS boundaries cross state lines<br>☐ State not identified |
|---|---|
| State | Alabama<br>Alaska<br>Arizona<br>Arkansas<br>California<br>Colorado<br>Connecticut<br>Delaware<br>District of Columbia<br>Florida |

**Person Variables**

| Age | from 0 to All |
|---|---|
| Sex | ☐ Male<br>☐ Female |
| Race | ☐ White<br>☐ Black/Negro<br>☐ American Indian<br>☐ Chinese<br>☐ Japanese<br>☐ Other<br>☐ Other, nec |
| Marital Status | ☐ Married, spouse present<br>☐ Married, spouse absent<br>☐ Separated<br>☐ Divorced<br>☐ Widowed<br>☐ Never married, single |

**FIGURE 7.** IPUMS data access system, 1995. Step 3.

---

[15]Gardner interview, Oct. 24, 2013; Sobek interview, Oct. 10, 2013.

Gardner initially estimated that it would take him "a couple of weeks" to transform his fantasy football program into the front end of a data extract system, but the task was harder than he anticipated. It took Gardner approximately six months before the inaugural launch of the website in November 1995, and thereafter, the system was in a continuous cycle of refinement for the next few years [31]. Unlike the early FTP dissemination system and the early IPUMS website, which required a download of entire census samples, the data extraction system allowed researchers to customize their data request to meet their research questions and computing capacity. Researchers followed prompts to select only those subpopulations and variables needed for their analysis and the output was a single data file containing multiple census years with identical custom record layouts [7], [31], [32]. A key element of the extraction system was the user interface that was developed using Perl and JavaScript. The expertise of SSRFC staff member John Easton was the key to untangling some logistical snarls in the computing process. The major design goal was "to make extractions easy while retaining maximum flexibility for users."[16] At the IPUMS landing page, users were guided through a step-by-step process to make selections about sample, variable, and case characteristics (see Figures 5–7). In 1996, the website was refined to accommodate fully automated online registration and dynamic pages generated from user selections and metadata. Between April 1995 (when the data extract system went live) and 1999, the system processed 10,000 data extract requests [32].

The convergence of increasingly fast and affordable personal computers; the expansion of access to UNIX workstations across universities; and the development of the Internet and the World Wide Web facilitated the dissemination of historical census data. The Internet allowed IPUMS to bypass physical dissemination of the data and the World Wide Web made it possible to customize and automate data extraction in an online environment.

A technological constraint parallel to the dissemination of historical census data was the creation and dissemination of integrated documentation. The sheer volume and inconsistency of the documentation across public use samples made its use unwieldy at best. Early IPUMS documentation, approximately 800 pages in length and contained in 61 separate files, was available in a downloadable Word for DOS 6.0 format.

A printed version of the documentation was available for purchase at $30 (see Figure 4). In 1996, Ruggles and Menard obtained funding to convert the documentation into hypertext format accessible on the World Wide Web. The hypertext documentation was integrated into the extract system, enabling researchers to make informed choices about the samples they were requesting [31].

## IPUMS 1996–1999

As the decade progressed, steadily expanding funding from both the NSF and the NICHD continued to support the creation of samples for additional censuses to extend IPUMS [24]. The growth of the census projects stretched the physical space and computing capacities where the data integration work took place.

During this period, the projects occupied three separate spaces on the West Bank of the University of Minnesota. As noted above, the Social History Research Lab began in space provided by the History Department. As the number of census project expanded, the department ran out of space. Desperate, and finding no on-campus solutions from the College of Liberal Arts, Ruggles recalled discovering that

"you can convert grants from on-campus to off-campus...If I had money, I could rent space. So, I arranged to rent space and then I just transformed three grants from being on-campus to being off-campus...Because we [had] just moved, we were able to shift money from indirect cost to direct costs. And so, we had more money available, and we were allowed to spend it on rent."[17]

The projects thus rented space in the Cedar-Riverside People's Center, a block from the campus. In addition to the census projects, the building housed a free medical clinic, a veterinary clinic, a gymnasium featuring nude volleyball, community theater instruction in theatrical combat, and a variety of "alternative" organizations. According to DEO Dianne Star, working at the People's Center was "interesting, to say the least," with "all sorts of people going in and out" of the building.[18] The project also found space in scattered on-campus locations. Data entry, data cleaning, processing, dissemination, and administrative functions were simultaneously carried out in buildings based on physical fit rather than function.

---

[16]Todd Gardner, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 7, 2021; Easton interview, Oct. 21, 2021.

[17]Steven Ruggles, interviewed by D. L. Magnuson, University of Minnesota, Jan. 9, 2014.
[18]Dianne Star, interviewed by D. L. Magnuson, University of Minnesota, Sep. 26, 2013.

The People's Center was a particularly challenging work environment for technical rather than social reasons. Initially, even basic phone service was problematic [13].[19] Dial-up modems were used to transfer data, but they were frustratingly slow and unreliable. A thinnet network connection was explored, but the $80,000 cost was prohibitive.[20] The solution was a wireless link connected by point-to-point microwave antennas. One antenna dish was located on the 12th floor of the Social Science Tower in the Political Science Department's computer lab; the other was positioned on the roof of the People's Center (0.2 mi/0.3 km away). Tom Lindsay, a graduate student liaison between SSRFC and the Social History Research Lab, had a table saw in his garage and built the frame to support the antenna mounted on the People's Center. A six-foot copper spike was buried next to the building with a copper wire run up the side of the building to ground the antenna. The Ethernet connection was threaded down through a skylight into the workspace.[21]

With the success and enthusiastic response by social scientists to the 1995 release of IPUMS, federal funding was obtained to launch a "comprehensive revision" of the database in 1998. Numerous variables were added, coding schemes were "improved and rationalized," and "virtually" all missing and illegible data were allocated through logical editing and "hot deck" probabilistic editing procedures. Furthermore, preliminary samples of 1860 and 1870 and an "expanded and improved" version of the 1920 sample were integrated into the IPUMS database. The extent of available documentation exploded from an 800-page guide to 3000 pages in a five-volume set [33]. As computing capacity increased and storage costs decreased across the decade of the 1990s, what once seemed out of reach was becoming normal.

## CONCLUSION

The IPUMS model proved highly scalable. In 1999, IPUMS expanded to incorporate data from censuses and surveys covering 102 countries, providing information on over a billion additional individuals. The North Atlantic Population Project provided access to international historical census data, including full-count census data from six countries. IPUMS now covers U.S. and international labor force surveys, time-use surveys, health microdata, and spatial population data.

IPUMS is now the world's largest population database, providing information describing more than three billion persons as well as billions of data points describing the characteristics of places. The size of the data collection is currently about four terabytes. When the project was proposed in 1991, much disk storage would have cost about $19.5 million in 2021 dollars. A decade earlier in 1981, the same storage would have cost some $2.6 billion, based on the cheapest prices for hard disk space advertised in *Byte Magazine*. By 2001, the cost of 4 TB of the disk had declined to less than $14,000, and today (in 2022) one can buy it for $69.99.

A convergence of technological advances over the past four decades made IPUMS possible. Throughout the 1990s, IPUMS pushed the limits of what was possible with the available computing resources. This involved improvising workarounds of the constraints of data storage, processing, and networking. IPUMS developed key innovations around structured metadata for data integration and web-based data access software. These innovations made large-scale data integration feasible and allowed users with limited computer capabilities to take advantage of an enormous (for the time) data resource.

IPUMS was initially conceived primarily as a resource for historical demography. The audience is now much broader, with over 200,000 registered users from dozens of academic disciplines. The largest contingent of users, accounting for some 40% of the total, are economists, followed by sociologists, demographers, geographers, and public health researchers. The availability of historical and contemporary data in consistent format has contributed to a dramatic increase in qualitative historical research across multiple disciplines [25]. Google Scholar lists 26,700 books, articles, working papers, and dissertations citing IPUMS data, and a new paper appears approximately every 3 h. Beyond the academy, IPUMS has found substantial audiences in policy, planning, and news organizations. By democratizing data access, reducing redundant effort, and enabling investigators to construct customized demographic datasets, IPUMS has opened new research opportunities and stimulated new discoveries.

---

[19]Dianne Star, interviewed by D. L. Magnuson, University of Minnesota, Sep. 26, 2013.
[20]Easton interview, Oct. 21, 2021; Thomas Lindsay, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Nov. 1, 2021.
[21]Lindsay interview, Nov. 1, 2021; Easton interview, Oct. 21, 2021; William Block, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 27, 2021.

## BIBLIOGRAPHY

[1] A. F. Anderson and L. Neidert, "Dissemination of data: Electronic produces," in *Encyclopedia of the U.S. Census: From the Constitution to the American Community Survey*, M. J. Anderson, C. F. Citro, and J. J. Salvo, Eds., Washington, D.C., USA: CQ Press, 2012, pp. 182–188.

[2] A. G. Bogue, "The historian and social science data archives in the United States," *Hum. Factors*, vol. 44, no. 4, pp. 976–987, 1976.

[3] J. Cushing and J. Ortuzar, "ISSA: An integrated system for survey analysis," *Population Index*, vol. 54, no. 3, p. 426, 1987.

[4] K. Dillard, "Farming in the shadow of the cities: The not-so-rural history of rose township farmers, 1850-1900," *Ramsey County Hist.*, vol. 20, no. 3, pp. 3–19, 1985.

[5] EFE, "Lab in Switzerland celebrates invention of the World Wide Web 30 years ago." Accessed: Aug. 1, 2022. [Online]. Available: www.efe.com/efe/english/technology/lab-in-Switzerland-celebrates-invention-of-the-world-wide-web-30-years-ago/50000267-3921488

[6] T. Gardner, "Software development," *Historical Methods*, vol. 28, no. 1, pp. 59–62, Winter 1995.

[7] T. Gardner, S. Ruggles, and M. Sobek, "IPUMS data extraction system," *Historical Methods*, vol. 32, no. 3, pp. 199–124, Summer 1999.

[8] S. N. Graham, *1900 Public Use Sample User's Handbook*. Seattle, WA, USA: Center Stud. Demography Ecol., Univ. Washington, 1980.

[9] P. K. Hall et al., "IPUMS metadata: Documenting 150 years of census microdata," *Historical Methods*, vol. 32, no. 3, pp. 111–119, Spring 1999.

[10] R. Jenkins, *Procedural History of the 1940 Census of Population and Housing*. Drive Madison, WI, USA: Center Demography Ecol., Univ. Wisconsin-Madison, 1983.

[11] D. Kallgren, "Project budget line-equipment record," ISRDI Archive, Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep. 12/1/1989-5/18/1990, 1880.

[12] D. L. Magnuson, "Curating our social science infrastructure: The MPC/IPUMS institutional history as a case study," presented at the *Social Sci. Hist. Assoc.*, Baltimore, MD, USA, Nov. 2015.

[13] D. L. Magnuson, "The spatial evolution of the MPC," Oct. 28, 2016. [Online]. Available: https://blog.popdata.org/the_spatial-evolution-of-the-mpc/

[14] D. L. Magnuson, "The evolution of our physical space," May 23, 2018. [Online]. Available: https://blog.popdata.org/physicalspace/

[15] R. McCaa, "Calidad, clase, and endogamy in colonial Mexico: The case of Parral, 1788-1790," *Hispanic Amer. Historical Rev.*, vol. 64, no. 3, pp. 477–502, Aug. 1984.

[16] R. McCaa, "Microcomputer software designs for historians: Word processing, filing and data entry programs," *Historical Methods*, vol. 17, no. 2, pp. 68–74, Spring 1984.

[17] T. J. Misa, *Digital State: The Story of Minnesota's Computing Industry*. Minneapolis, MI, USA: Univ. Minnesota Press, 2013.

[18] "Mosaic—The first global web browser." Accessed: Aug. 2, 2022. [Online]. Available: www.livinginternet.com/w/wi_mosaic.htm

[19] S. Ruggles, *Prolonged Connections: The Rise of the Extended Family in Nineteenth-Century England and America*. Madison, WI, USA: Univ. Wisconsin Press, 1987.

[20] S. Ruggles, "Integration of the public use samples of the U.S. census," in *Proc. Amer. Stat. Assoc., Social Statist. Sect.*, 1991, pp. 365–370.

[21] S. Ruggles, "The U.S. public use census microdata files as a source for the study of long-term social change," *IASSIST Quart.*, vol. 15, no. 2, Summer 1991, Art. no. 20, doi: 10.29173/iq703.

[22] S. Ruggles, "Integrated public use microdata series," NSF, Alexandria, VA, USA, Tech. Rep. SES-9118299, 1992–1995.

[23] S. Ruggles, "Public use microdata sample of the 1920 census," NICHD-DBSB, Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep. R01 HD29015, 1993–1998.

[24] S. Ruggles, "The Minnesota population center data integration projects: Challenges of harmonizing census microdata across time and place," in *Proc. Amer. Stat. Assoc., Government Statist. Sect.*, 2005, pp. 1405–1415.

[25] S. Ruggles, "The revival of quantification: Reflections on old new histories," *Social Sci. Hist.*, vol. 45, pp. 1–25, Spring 2021.

[26] S. Ruggles and D. L. Magnuson, "Census technology, politics, and institutional change, 1790-2020," *J. Amer. Hist.*, vol. 107, no. 1, pp. 19–51, Jun. 2020.

[27] S. Ruggles and R. R. Menard, "A public use sample of the 1880 U.S. census of population," *Historical Methods*, vol. 23, no. 3, pp. 104–124, Summer 1990.

[28] S. Ruggles and R. Menard, "Public use microdata sample of the 1850 census," Sociol. Division, NSF, Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep. SBR9210903, 1992–1994.

[29] S. Ruggles and R. R. Menard, "Public use sample of the 1880 census," NICHD-DBSB, Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep. R01 HD25839, 1989–1993.

[30] S. Ruggles et al., *Public Use Microdata Sample of the 1880 United States Census of Population: User's Guide and Technical Documentation*. Minneapolis, MO, USA: Social Sci. Hist. Res. Lab., Univ. Minnesota, 1994.

[31] S. Ruggles, M. Sobek, and T. Gardner, "Distributing large historical census samples on the internet," *Hist. Comput.*, vol. 9, pp. 145–159, 1996.

[32] S. Ruggles, M. Sobek, and T. Gardner, "Disseminating historical census data on the World Wide Web," *IASSIST Quart.*, vol. 20, no. 3, 1996, Art. no. 4, doi: 10.29173/iq68.

[33] M. Sobek and S. Ruggles, "The IPUMS project: An update," *Historical Methods*, vol. 32, no. 3, pp. 102–110, Summer 1999.

[34] M. A. Strong, S. H. Preston, and M. C. Hereward, "An introduction to the public use sample of the 1910 U.S. census of population," *Historical Methods*, vol. 30, no. 2, pp. 54–56, Apr. 1989.

[35] University of Minnesota, *Minnesota Project Woksape: Innovation, Learning, Wisdom. IBM Advanced Education Project*. Minneapolis, MN, USA: Univ. Minnesota, 1989. [Online]. Available: https://conservancy.umn.edu/handle/11299/207708

[36] University of Minnesota, "West bank computer center," *West Bank Buffer*, Nov. 1973. [Online]. Available: https://conservancy.umn.edu/handle/11299/162893

[37] University of Minnesota, "The president's report, 1964-1966," 1966. [Online]. Available: https://hdl.handle.net/11299/102308

[38] U.S. Bureau of the Census, Census of Population, *1940: Public Use Sample Technical Documentation*. Washington, D.C., USA: U.S. GPO, 1984.

[39] U.S. Bureau of the Census, Census of Population, *1950: Public Use Sample Technical Documentation*. Washington, D.C., USA: U.S. GPO, 1984.