

Project Description

Objectives

This project will create and freely disseminate a database incorporating all available aggregate census information for the United States between 1790 and 2000. Enormous quantities of machine-readable aggregate census data exist, but they are inaccessible. The project entails three complementary tasks. First, we will collect and enrich data and metadata that will support broad-based investigations in the social and behavioral sciences. Second, we will incorporate these data into a Geographic Information Systems framework. Finally, we will create a system incorporating innovative capabilities for worldwide web-based access to both census data and the metadata needed to interpret it.

This is basic infrastructure for the social sciences and it is urgently needed. Shockingly, there are no national electronic maps for small areas in the United States covering the period before 1980, and there are no high-quality maps available for the period before 1990. This makes the analysis of geographic change in the American population exceedingly difficult. When researchers do assess change, they must develop their own maps at great expense. Not surprisingly, the number of such studies is small and their geographic and chronological scope is highly limited (Duncan 1957, Leiberson 1961, Denton and Massey 1991, Massey and Eggers 1990; Massey and Denton 1988, Alba et al. 1995, Logan et al. 1996).

Even raw machine-readable counts of the population characteristics of local areas are inaccessible except to a few experts. Save for a fraction of the most recent data, existing aggregate data are available only in obsolete nonstandard formats accompanied by paper documentation. Census information for the period before 1940 is especially inaccessible, since most of it has never been converted to machine-readable form and exists only in paper form housed at various archives around the country. The 1960 and 1970 tract-level data are available only in a special early-1970s compressed format developed by DUALabs, Inc., a private census reseller that has been out of business for twenty-five years. Many scholars are not even aware that small-area machine-readable census data survive for these years. Most of the 1970 census tract data were thought to be lost until a full set was discovered at the Census Bureau in 1997. Similarly, a valuable set of tract data spanning the period 1940 through 1970 that was laboriously keypunched at the University of Chicago a quarter-century ago was recovered only in late 1999 (<http://www.nara.gov/nara/electronic/bogue.html>).

This project will gather together all surviving census data, put them into consistent format, develop comprehensive standardized machine-readable documentation, create high-precision historical electronic boundary files describing census tracts, civil divisions, and counties, and develop innovative web-based tools for disseminating both microdata and metadata over the Internet.

The proposed database is very large. It will include roughly the same quantity of data as presently exist in the entire archive of the Inter-university Consortium for Political and Social Research—the world's largest social science data archive. This scale of infrastructure would have been unthinkable just a few years ago. Five recent technological innovations make this project feasible:

1. The decline in the cost of data storage during the past five years makes it possible to maintain the entire body of machine-readable census data online.
2. The development of the Internet has slashed the cost of worldwide dissemination.
3. The development of Geographic Information Systems (GIS) technology has provided the tools to describe and display changes in census geography. GIS methods and concepts enable us to create consistent machine-readable census geography across time and thereby allow coherent chronological and comparative analysis of small-area census data.
4. The development of the Data Documentation Initiative (DDI) international metadata standard—finalized in March 2000—gives us the essential tool for automatic processing of census documentation. Without the DDI, it would be far more expensive to manage this vast collection of data and documentation.
5. The development of advanced web-based data extraction tools at the Minnesota Machine Readable Data Center, the Minnesota Population Center and elsewhere has made it possible to simplify access to complex data structures. This means students, policy analysts, journalists and academic researchers will not need specialized training to make use of the data.

This proposal capitalizes on all these developments. They will allow us to create a database of unprecedented size and power and to make this resource readily accessible for analysis on desktop computers.

Background and Significance

The census is the primary source of statistical information about growth and change of the American population since 1790. Aggregate data tables, in print or electronic format, are the principal means of describing the characteristics of states, metropolitan areas, cities, counties, villages and neighborhoods. Approximately 670 gigabytes of U.S. Census summary data covering the period 1790 through 2000 exist or are in preparation, but they are scattered across dozens of archives and are stored in incompatible formats on magnetic tape, CD-ROMs, or on paper.

Source data. This project will incorporate all surviving machine-readable aggregate census data for the United States and will add new data transcribed from printed and manuscript sources. Table 1 lists the principal datasets. They describe the characteristics of states and counties, census tracts, cities, minor civil divisions, census blocks and zip codes.

The most important source for state and county data before 1950 is “Historical, Demographic, Economic, and Social Data: The United States, 1790-1970,” a dataset created thirty years ago by the Inter-university Consortium for Political and Social Research with funding from the National Science Foundation (ICPSR study 0003; <http://www.icpsr.umich.edu/cgi/ab/prl?file=3>). This dataset includes the bulk of published nineteenth and early

Table 1. Summary of principal datasets to be included in the NHGIS

Dates	Dataset Description	Size MB	Variable Count	Available On-Line
1790-1970	County and state (ICPSR 0003 and additional files)	253	3,508	partially
1790-1990	County and state election returns (ICPSR 1, 2, 13)	182	2,867	no
1790-1960	Municipal data (Haines files; size estimated)	160	950	no
1944-2000	County and city data books	66	966	no
1974-2000	County business patterns	3,344	65	no
1947-2002	Economic censuses	3,400	1,000	no
1949-2002	Agricultural censuses (Gutmann)	2,000	4,000	no
1940-1970	Census tract data (Bogue files)	211	2,478	no
1910-1950	Supplemental tract data (Beveridge files)	150	2,658	no
1960	Census tract level data	246	1,073	no
1970	Census small area data			
	Count 1	275	447	no
	Count 2	710	7,000	no
	Count 4	3,224	4,300	no
1980	Census summary files			
	Summary tape file 1 A-H	10,193	321	no
	Summary tape file 2 A-C	12,697	2,292	no
	Summary tape file 3 A-D	7,328	1,126	no
	Summary tape file 4 A-C	22,713	1,500	no
	Summary tape file 5	5000	100,000	no
	Equal opportunity employment file	415	1,100	no
	Journey-to-work file	172	88	no
1990	Census summary files			
	Summary tape file 1 A-D	60,476	1,050	partially
	Summary tape file 2 A-C	33,966	2,187	no
	Summary tape file 3 A-D	33,625	3,225	yes
	Summary tape file 4 A-C	157,826	22,040	no
	Special summary tape files: 1-22	15,689	44,600	no
	File 420 Place of work twenty destinations	541	406	no
	File S-5 Workers by residence by workplace	22	19	no
	Equal opportunity employment file	909	13,135	no
2000	Census summary files (size estimated)	300,000	80,000	partially
TOTALS		671,293	304,401	

twentieth-century state- and county-level statistics from the censuses of population, agriculture, manufacturing and religion. Unfortunately, the file is incomplete and is plagued by numerous data-entry errors. We will correct the errors and augment the dataset with additional information from published and machine-readable sources.¹ To allow analysis of political change, we will also include the ICPSR county-level election return studies, which cover the period 1790 through 1990. For the period since 1950, we will supplement the ICPSR files with machine-readable data from the county data books, the economic and agricultural censuses and the county business patterns data files.

The most important statistical unit below the county level is the census tract. The tract system was first applied to eight cities for the 1910 census. By 1930, nineteen cities were tracted, and from 1940 onward most metropolitan areas and other densely populated counties were enumerated and tabulated by tract. For the early period—from 1910 to 1930—census tract data exist only in manuscript form except for New York City, which was digitized under the direction of Andrew Beveridge. Elizabeth Mullen Bogue digitized most of the 1940 tract data and about sixty percent of the 1950 data thirty years ago. These data files, long thought to be lost, were recovered in 1999. Professor Beveridge will correct the Bogue data and will convert all remaining tract data from 1910 to 1960 into machine-readable form. The 1960 and 1970 tract data—also recently recovered—survive in machine-readable form but are presently stored in an obsolete format. Tract data for 1980 and 1990 exist in ASCII files on magnetic tape, and some files for 1990 are available on CD-ROM. The 2000 tract files will be distributed exclusively on CD-ROM and via the Internet.

In addition to the state, county and tract data, we will incorporate data on cities, minor civil divisions (MCDs), census blocks, zip codes, and other census designated places. Machine-readable data on cities and MCDs are available only for the period since 1950; Michael Haines will extend these series back into the nineteenth century by digitizing published census returns. Block and zip code data are available only for the period since 1980.

Data preservation and access. The NHGIS database will address one of the most perplexing problems currently facing depository librarians, that of preserving and maintaining functional access to government data. The National Archives preserves much of the machine-readable census data and documentation, but no institution maintains the software that was designed to provide functional access.

Many summary data files produced from 1960 through 1980 came without search software. Software tools that did exist, such as the Census Software Package (CENSPAC), were dependent on particular hardware and software that are no longer maintained. Thus, today's researchers actually have no functional access to machine-readable aggregate data for the 1960, 1970 and 1980 censuses. Virtually no small-area data for the period before 1990 are available on the Internet.

There is every reason to believe that the same basic problems of hardware and software dependence, obsolescence and loss of functional access will arise with the 1990 and 2000 census data released on CD-ROM. Indeed, maintaining functional access to data products is in some respects a more serious problem for the 1990 and 2000 censuses than earlier census years because the Bureau has abandoned printed publication of small area data. The National Archives is preserving raw census tables in machine-readable form, but it cannot maintain the operating-system-dependent software needed to process, extract, and analyze the information. The majority of data users lack the expertise to look up particular items in the raw data files, so long-term access is endangered.

For the time being, selected 1990 data are accessible from multiple Internet sites with basic look-up and subsetting functions, but the great bulk of the 1990 census data are still unavailable. Even within the Census Bureau, access to some of the most useful 1990 tract data is a cumbersome procedure involving specialized software, auxiliary files, templates and highly skilled personnel.

Chronological analysis multiplies the complexity of data access. Since software and file formats differ in every census year, it is nearly impossible to assess change over time. A generic search and extraction engine, however, could present the contents of each data file intelligently without the need for customization. To create such an engine, we need a method for encoding documentation in a form that allows automatic processing of the data. In other words, we need comprehensive machine-understandable documentation.

¹ We have obtained approximately 65 state and county files prepared by eight different scholars that will supplement the ICPSR data, but several important gaps remain which will have to be entered directly from published sources. The raw data from ICPSR study 0003 are available for viewing on-line (<http://fisher.lib.virginia.edu/census>), but they cannot be downloaded and no mapping or statistical facilities are available.

The solution is within reach. In May 1995 the executive director of the Inter-university Consortium for Political and Social Research (ICPSR) established a Data Documentation Initiative (DDI) committee to develop a new standardized structure for social science data documentation (<http://www.icpsr.umich.edu/DDI/codebook.html>). The international committee represented a range of stakeholders in social science data dissemination, including the Census Bureau, the Bureau of Labor Statistics and the national data archives of Great Britain, Norway, and Canada. The work was funded by National Science Foundation grants, ICPSR membership dues and thousands of contributed hours by participants. The results of this work, a document type definition in the Extensible Markup Language (XML), was published in March 2000. The machine-understandable structure of the DDI allows for automated processing by data access software. It is a non-proprietary, hardware independent, neutral standard that preserves the content and relational structure of the full documentation.

Because the DDI is being adopted by most of the world's leading data archives and statistical agencies, we can be confident that it will not become obsolete in the foreseeable future. As part of an NSF Infrastructure grant, ICPSR is now converting the documentation for its holdings into DDI-compliant form. Eventually, of course, new metadata standards will emerge. The enormous base of DDI-compliant documentation that will soon exist ensures that software will be available to migrate the entire documentation system smoothly with minimal human intervention.

Geographic Information Systems. The DDI has a critical limitation: it cannot express the basic geographic units used by the Census Bureau. The census data describe small geographic areas, and the boundaries of those areas were modified in each census year. In the past, researchers used small-area census data in conjunction with paper census tract maps, so they could visually identify the places concerned. This approach is unwieldy for analyzing change across time or across many geographic areas.² To open the body of census data to chronological and spatial analysis, it is essential that we provide machine-readable descriptions of the places covered by the statistics.

In recent years, technological change has revolutionized the field of population geography. We now have powerful computer-based methods for the acquisition, storage, analysis and display of spatial data. These Geographic Information Systems (GISs) have significantly broadened the scope of questions that can be answered with geospatial data and have popularized the use of mapping techniques for the display of spatial information.

At the core of GIS technology are electronic boundary files that describe the spatial dimensions of each geographic entity. High-quality machine-readable census boundary files for small areas exist only for the 1990 and 2000 censuses. Low-definition boundary files for 1980 census tracts also exist. We propose to create high-quality census tract maps back to 1910, state and county maps back to 1790, and minor civil division maps where feasible. The availability of high-quality boundaries for the key statistical areas will allow us to reconcile changes in census geography. This in turn will make it possible to provide researchers with estimates of changing population characteristics and distribution under constant definitions of census geography.

Web-based dissemination. The key to making census summary files broadly accessible is the development of a powerful and flexible web-based data access system. We have been working on methods of electronic dissemination for social science data and documentation since 1993, and have developed the most powerful and widely used tool for access to census microdata (<http://www.ipums.org>). We plan to develop the NHGIS data access system according to the same principles: ease of use will be paramount and full documentation will be seamlessly integrated with the data extraction functions.

The system will consist of a set of tools for navigating the mass of documentation, selecting variables, and creating formatted tables and thematic maps. In addition, users will be able to extract downloadable datasets and boundary files in a variety of formats. Users will create customized subsets of both data and documentation tailored to their particular research questions. The system will be intelligent enough to allow use by novices but flexible enough to meet the needs of advanced analysts. For example, users will have three options for geographic case selection: a clickable map interface, a scrollable structured list, or a search utility that will return a list of geographic units that match any part of a given place name. We will provide a variety of tools to allow users to locate appropriate variables and will provide basic statistical and arithmetic functions to allow users to calculate percentages, means and ratios using any desired denominator.

² It is now possible—although difficult—to map changes across recent census years in areas where few tract changes exist (Denton and Massey 1991, Alba et al. 1995). Studies with longer chronological scope are so time consuming and tedious that they are seldom undertaken.

Research applications. The ready availability of aggregate census data in a GIS framework will have revolutionary consequences for a broad range of research problems. Among key substantive areas are residential segregation; the decline and renaissance of central cities; immigrant and ethnic settlement patterns; suburbanization and urban sprawl; rural depopulation and agricultural consolidation; the identification of concentrated poverty; causes and levels of change in ecosystems; transportation; public health and epidemiological studies; the transformation of electoral politics; geographic criminal justice studies; environmental justice; and multilevel analysis integrating aggregate census data and micro-level data. The potential list of topics that can be addressed with these data is far too long to discuss within the space constraints of this proposal. The following paragraphs give a few examples to illustrate the exciting possibilities opened up by the NHGIS.

Residential segregation studies and urban change. Although social scientists have developed a multitude of segregation indices over the past ten years, nearly all work applies such indices to a single moment in time. Moreover, most studies focus on simple indices of segregation rather than assessing changing residence patterns (Taeuber and Taeuber, 1957, Massey and Denton, 1998, Farley and Frey, 1996). Through painstaking reconstruction of census geography, Alba et al. (1995) were able to assess factors that predict changes in specific neighborhoods over a twenty-year period (see also Wyly and Hammel 1998). Their study required the sort of data for New York that we propose to create for the entire United States. Moreover, by linking microdata to tract-level data, analysts can assess how neighborhood change is associated with individual behavior (Scott and Crowder 1997). The NHGIS promises to revolutionize the study of urban racial and ethnic change by allowing both long-run analysis of change and comparison of cities within regions or across the nation.

Public health and epidemiological studies. The field of epidemiology is rapidly turning to spatial solutions both to trace the path of diseases and to determine their impact on subpopulations. A plethora of studies now uses geographical analysis of census data for public health analysis (e.g., Becker et al. 1998; Leclere, Rogers and Peters 1998; Sucoff and Upchurch 1998; Miles-Doan 1998; Latkin, Glass and Duncan 1998; Liu, Deapen and Bernstein 1998; Sayegh et al. 1999; Arbes et al. 1999; Nuorti et al. 2000). Spatial-temporal epidemiological investigations, however, remain difficult. Some historical data on the incidence of disease by county and census tract are already available (e.g. <http://www.nci.nih.gov/atlas>; Lang and Polansky 1994), but epidemiologists usually have not had access to appropriate historical small area population counts needed to calculate morbidity rates for population subgroups or to control for historical variation in age composition. The NHGIS database will expand the potential for sequential analyses of public health. Changes in patterns of cancer, for example, can be related to demographic shifts as well as to changes in radiation, toxic waste sources and other environmental characteristics. Increasingly, researchers have argued that a historical perspective is critical for analysis of the impact of toxic sites on disease incidence (Pulido 2000).

Ecological change and policy. Along with aggregate data from the population, agricultural, and economic censuses, the NHGIS will include information on land use, climate and physical geography. Together, these data will provide rich comparative information and document change at the community level in population density and distribution, economic development, resource availability and extraction, and land use (Gutmann 2000). Moreover, investigators will be able to link these data with biological and ecological information on habitat classification, pollution and species distribution (Woodruff et al. 1999, Mielke et al. 1999). These complementary small-area data will promote collaboration among social and natural scientists and policymakers for research, natural resource management, land use planning, and education. Thus, the NHGIS will help to address broader goals such as balancing biodiversity conservation and human development, and detecting causes and levels of historical change in ecosystems.

The NHGIS will not only be the world's largest social science database, it will also be one of the most powerful. By creating the infrastructure to access this vast collection, we will allow social scientists for the first time to address simultaneously the broad sweep of time and the detail of spatial organization. This power to analyze variation in human behavior simultaneously across both time and space will stimulate innovation across fields ranging from history to epidemiology.

Educational and public applications. Despite the value of the NHGIS to social science research, use of the database by educators, journalists, planners and market analysts will be even greater. The opportunities for classroom use are especially exciting. During the 1980s, David Miller of Carnegie-Mellon University developed a computer-mapping program for use on Unix workstations called "The Great American History Machine." It allowed students to create thematic maps based on the state and county data files produced by ICPSR, and to explore the relationships among variables. Though it taxed the limits of available technology and was somewhat difficult to

use, the software quickly became the central focus of at least a dozen courses in social history, political science and social demography (Miller and Modell 1988). Technological change rendered the Great American History Machine obsolete. The NHGIS will have all the same capabilities but will incorporate data on a far broader range of subjects, will have higher quality data and higher-definition boundary files, will be easier to use, and will be accessible at no cost to all instructors who have access to networked computers. It will also offer finer geographic detail, allowing tract mapping as well as county mapping. By focusing on their own neighborhoods, students will be able to apply theoretical concepts and statistical methods to familiar terrain. All this means that the NHGIS will find an audience in secondary education as well as for college courses in history, sociology, political science, public health, ecology, and statistics.

We anticipate an equally enthusiastic response among journalists, community planners and the private sector. News organizations are avid consumers of local statistics and thematic maps. For example, the *New York Times* has published 144 articles using historical census research on New York carried out by Andrew Beveridge. Reducing the complexity of data access and increasing the available detail of census information will improve the ability of planners to evaluate community needs. Local area statistics are an essential tool for decisions concerning education, transportation, care of the aged, and a host of other community planning issues. Finally, we expect extensive use of the NHGIS by the private sector. Large firms have long made use of small-area census data for market research. By reducing the cost of access to these data, the NHGIS will bring the same resources to smaller companies. Because the NHGIS offers the opportunity to assess demographic and economic trends at the local level, all users will be able to move beyond static analysis and better prepare for the future. In short, this project not only provides fundamental infrastructure for the social sciences, but also promises to expand the use of such material by the public.

Institutional context. Despite the critical importance of this work, no federal agency is able to undertake it. The Census Bureau, preoccupied with the need for current data, is sympathetic but cannot address the problem of access to past censuses. The National Archives has neither the resources nor the expertise to carry out the task. This is a big job and it is of critical importance.

The Minnesota Population Center (<http://www.pop.umn.edu>) is uniquely equipped to take on this challenge. We have extensive experience with aggregate census data. Since its inception in 1990, our Machine-Readable Data Center (MRDC) (<http://www.lib.umn.edu/mrdc>) has focussed on problems of access to aggregate census data. The MRDC was a beta-test site for the DDI international metadata standard underlying this project. Wendy Treadwell, coordinator of the center, designed the extensions of the DDI necessary for encoding aggregate datasets and is now completing a pilot project to develop a web-based display and extraction system for aggregate public data (<http://www.socsci.umn.edu/PDAS>).

We also have exceptional cartographic resources. Our Cartography Laboratory, housed in the Department of Geography (<http://www.geog.umn.edu>) is one of the most advanced in the nation and has extensive experience in mapping census tracts and minor civil divisions. Additionally, the Automated Cartographic Information Center (<http://map.lib.umn.edu>) within the Borchert Map Library maintains a state-of-the-art GIS facility including digital copies of all Census Bureau electronic maps. The Geography Department is also home to the oldest and largest Master of Geographic Information Science (MGIS) program in the country. This program will provide us with a highly skilled pool of labor for the cartographic component of the project. The proposed project will also pay important educational dividends: we will integrate the NHGIS into the MGIS curriculum and anticipate that the project will provide several cohorts of graduate students with invaluable practical experience.

We have successfully completed several large social science infrastructure projects. Most notably, we created the Integrated Public Use Microdata Series (IPUMS), a coherent series of individual-level U.S. census data drawn from thirteen census years between 1850 and 1990 (Ruggles and Sobek 1998; <http://www.ipums.org>). By putting all U.S. census microdata samples in a compatible format with consistent variable codes and integrating their documentation, the IPUMS greatly simplifies the use of data from multiple census years. Because of the IPUMS project, the Minnesota Population Center is already one of the largest distributors of census data in the world. The present proposal—which concerns aggregate census data rather than data on samples of individuals—will complement and enhance the IPUMS work. Indeed, many queries received by the IPUMS could be better addressed using aggregate data. The IPUMS is sample data, and information on small areas is either not available or is statistically unreliable. Yet, much of our IPUMS work—such as our on-line documentation of census procedures—will directly enhance the new project. Moreover, our experience with large-scale data dissemination and complex project management will serve us well.

When we released the IPUMS database in 1995, many researchers started using it simply as a means to access the 1990 census microdata. The IPUMS offered several advantages over the Census Bureau version of the 1990 microdata; we fixed numerous errors, provided better documentation, developed superior constructed variables and provided a more user-friendly data access system. In a brief period, the IPUMS became an essential source for research in history, sociology and economics: over a hundred articles, four books and twenty-five dissertations have used the IPUMS in just the past four years. But even if users initially adopted the IPUMS as a means of accessing the 1990 census, they quickly began to exploit the power of the IPUMS to carry out analyses of change over time. The overwhelming majority of IPUMS studies now make use of multiple census years. Indeed, the IPUMS is responsible for a dramatic increase in the total number of quantitative studies of long-run social and economic change. We believe that the NHGIS—which has broader chronological scope than the IPUMS, a wider range of subject coverage and far greater opportunities for geographical analysis—has an even greater potential to affect the course of research in the social sciences.

Plan of Work

We have subdivided the project into three closely interrelated work components. The *data and documentation* component will collect the data and documentation and convert them to a form suitable for redistribution. The *mapping* component will create historical electronic maps describing census geography. The *data access* component will develop extraction, browsing and mapping software for data and metadata. Each of these components depends on the others. Without developing electronic maps, it would be impossible to make consistent statistical comparisons across census years; without the work on metadata and data, there would be no statistics to compare; and without the development of new data access tools, the data, metadata, and electronic maps would remain largely inaccessible. The following sections detail the methods and procedures of each component in turn.

1. Data and Documentation

Data and documentation lie at the heart of the project. We will acquire and clean the census data and documentation, harmonize its format, and create DDI-compliant codebooks for each file. A central task is to convert all metadata pertaining to the aggregate U.S. censuses into DDI form. This step will allow us to rationalize the structure of the data files automatically and to implement a single extraction and browsing tool for the entire collection of data. In addition, the data and metadata component will fill in gaps in the existing machine-readable data series with the assistance of subcontractors at the University of Texas-Austin, Colgate University, and Queens College-CUNY. Data and documentation account for approximately forty percent of the total cost of the NHGIS project.

Data acquisition and cleaning. We intend to include all surviving aggregate data from the population, economic, and agricultural censuses. The database will also incorporate several other small data files, including selected voting statistics, land-use data and information on physical geography. We have already obtained documentation for all principal data files. Most existing machine-readable census data are either already held by the Machine-Readable Data Center or will be obtained at no cost from other State Data Center agencies and archives around the country. We will obtain files distributed through the Federal Depository Libraries Program from the University of Minnesota Government Publications Library. There are a few files held solely by the National Archives and Records Administration, which we must purchase.

The only machine-readable tract-level data available before 1960 are the Bogue files, which extend back to 1940. Unfortunately, these files do not cover all tracted cities and parts of them are unreadable. In addition, there exist printed tract data for nineteen cities in 1930 and eight cities in 1910 and 1920. We will contract with Andrew Beveridge of Queens College, City University of New York, to enter all missing tract data. Beveridge has already done this work for New York City. We plan to include city and county data back to 1790; we will contract with Michael Haines of Colgate University, the leading expert in the field, to fill gaps in those series. Finally, we will contract with Myron Gutmann of the University of Texas to assist us in developing a county-level series of data from the agricultural census for the period since 1850.

We will convert all files stored in EBCDIC or DUALabs compressed format to column-format ASCII. We will review all published user notes related to the datasets and assure that all identified corrections have been made in the data. To ensure data integrity, we will verify table sums against published totals within the same geographic area. We will then evaluate records with mismatched entries for appropriate corrections.

Creation of XML DDI codebooks. Source codebooks for aggregate census files are currently in three different formats: print, electronic images of print, or ASCII text files of the original print file. In some cases, only the data dictionary portion of the original documentation is available in ASCII format. We will create scanned images of all source documents for archival and reference purposes. Although some documentation exists as scanned images, these files are often unusable because of poor resolution. Creating archival photo-quality images (600 dpi to bi-tonal TIFF type five) is the goal. We will then convert these files to PDF documents, which will allow us to make them available to users on the Web.

The PDF codebooks will also be our primary source of information for the creation of DDI-compliant metadata. Each piece of information from the codebooks will be labeled with a “tag” that identifies the particular type of information, such as title, author, variable description or variable label. This mark-up will be carried out using a combination of customized XML authoring tools and standard editors.

Some census files, especially those from recent years, have either an ASCII text file of the data dictionary with good structural integrity or a data definition file for SPSS or SAS processing. The information contained in these files can be converted automatically to DDI format. We will write PERL scripts to map data into the appropriate elements wherever possible to reduce the amount of hand entry required, which will reduce data entry errors.

Accuracy of the electronic codebooks is essential. The data access system depends on the quality of the metadata, and without an accurate codebook, the results could be meaningless. We will use three verification methods to assure quality control. One or more of these methods will be used for each section of the codebook. First, we will carry out blind verification through double entry, particularly in textual areas. Second, we will create a template that duplicates the original print layout and visually compare the new documentation with the original paper documentation. Third, we will use the marked-up documentation to read the data and calculate check sums. If sums across variable categories and geographic entities are correct, this will verify that all variables within a matrix have been entered and that the start position, end position and width of each variable are consistent with one another and result in the specified record layout.

Harmonization. We will reorganize the files into an integrated format that approximates the organization of the Census 2000 summary files. We will harmonize geographic place names where geographic boundaries remain unchanged. Where geographic boundaries changed without a corresponding name change, we will assign new unique geographic identifiers. We will construct compatible variables across time where possible through summing, subtraction or other standard transformations. For example, wherever possible we will harmonize decade to decade changes in table universes and variable categories.

Preparation of ancillary documentation. Metadata are not confined to codebooks. We will bring together supplemental documentation to form a comprehensive collection of reference materials for the data user. These materials will provide full detail on geographic, occupational and industrial coding schemes, the construction of poverty indices and other derived variables over time, methodological papers, census questionnaire and product development reports, sample designs and sampling errors, procedural histories of each dataset, full documentation of error correction and other post-enumeration processing, and analyses of data quality. Much of this background material already exists as part of the on-line documentation of the IPUMS project at the Minnesota Population Center, but we plan to mark up these files to allow full integration with the NHGIS data access system.

We will prepare new documentation to address data harmonization, the data creation process for new data files, descriptions of incompatibility in variables or definitions among and within data series, the application of data to specific research issues, and discussion of linking applications and limitations. We will convert all documents to the appropriate XML format so we can link them to the data through the DDI structure. We will document our own procedures and data transformations as we proceed. Our codebook authoring tools allow us to capture the process of data selection, data processing and data file structure development. The goal is to limit the amount of double entry required to document the process, to create instructional material for coders and to create the final metadata.

2. Mapping

The need for computerized census maps is fundamental. To do any chronological analysis of the census, we need to know how geographic units were altered from one census to the next. The availability of maps will unlock the potential of the aggregate census data and will make possible whole new topics of historical geographic analysis. The mapping component of the project is substantial. We will scan approximately 7,000 source maps and create 630,000 cleaned and verified polygons. Cartographic work accounts for forty percent of the total budget.

The unit of analysis in aggregate census data is a geographic entity: the census block, tract, minor civil division, county division, county, city, metropolitan area or state. All these units—even states—can change from one census year to the next. Our primary cartographic emphases are census tracts and counties, the basic building blocks of the Census Bureau statistical system. The finest level of geography in the electronic maps will be the census tract. Census tracts usually have between 2,500 and 8,000 persons and are designed to be homogeneous with respect to population characteristics, economic status and living conditions.³ In addition to the tract maps, we will develop new high-precision county maps covering the entire country since 1790. We will also construct geographic entities that are aggregates of tracts or counties, such as metropolitan areas. County boundaries are comparatively stable—especially in comparison to urban areas, cities or census designated places. Tract boundaries are also relatively stable, by design, although tracts have frequently been subdivided or merged.

We have a secondary emphasis on county subdivisions. County subdivisions—minor civil divisions (MCDs), MCD-equivalents and census county divisions—form an intermediate layer between census tracts and counties. Of these, the census county divisions are the easiest to work with because they were designed to be aggregates of tracts in places where there were tracts. Census county divisions exist for twenty-one states, but they date only to the 1950 census. In the remaining states and in all states before 1950, MCDs and MCD-equivalents provide the primary sub-county areal units. The exact form of these units varies considerably around the country. They include incorporated places, towns, townships and other political units. For some areas—especially in earlier census years—we have been unable to locate appropriate base materials to create county subdivision boundary files. Therefore, we do not expect to be able to create boundaries for all county subdivisions in all census years, but we will do as much as is feasible.

Constructing census tract databases. High quality boundary files will be available for both the 1990 and 2000 censuses. Our task is to construct similar files for earlier years. We will begin by developing a clean set of boundaries for the 1990 and 2000 censuses. We will then work backward from census to census by undoing the boundary changes of each census year. This approach will allow us to use existing digital data as a basis for generating all base maps. In addition, it minimizes work needed for generating tracts for an earlier census year. The two most common geometric changes over time are the addition of new tracts and splitting of existing tracts; the most frequent editing operations, therefore, will be eliminating tracts and merging tracts by removing their common border. By reusing borders across different years we maximize the geographic correspondence between the different datasets and avoid the problem of map conflation.

We will use on-screen editing directly on digital files to minimize the high cost of using a graphics tablet digitizer. The primary dataset for each pre-1990 census will be a copy of the cleaned tracts for the succeeding census, and we will directly edit those files. Cartographic staff will scan paper tract maps for the appropriate year and rectify them to the cleaned tract files. These scanned images will then provide a base for editing the cleaned tract files.

The Census Bureau has created geocoded street and enumeration-unit files—the TIGER files—that allow spatial analysis and mapping. We will assemble the TIGER data into single layers, by county, and use them in the production of all pre-1990 tract datasets. The TIGER layer provides two essential functions. First it will be used as a general reference, especially for verifying tract borders by street name. Second, it will be the first choice for obtaining additional line work. For example, if a tract border changed over time from a railroad line to a street, and if the railroad still exists in the TIGER file, we will copy the appropriate line from the TIGER layer directly into the tract base map.

In cases where new borders must be added that are not found in the TIGER data, we will draw from a variety of additional ancillary data including maps in the John R. Borchert Map Library at the University of Minnesota. In some instances, we will have to resort to materials held elsewhere, including other university libraries, the Library of Congress and the National Archives.

The tract databases will contain corresponding attribute data describing each tract's history. For example, we will be able to query a particular 1990 census tract and find out that it existed with identical borders for 1980 and 1970, but that the 1970 tract had a different tract identifier (ID). Likewise, we will be able to query a tract that was newly-created in 1950, determine that it existed with identical borders in 1960, that it was split into two tracts for 1970, and

³ Block-level maps will be included for the 1990 and 2000 censuses, but we will not construct them for earlier years because of high cost and because no pre-1980 census detail is available at the block level.

obtain its sibling tract IDs for 1970 and subsequent censuses. These attribute data will be an essential building block of the data access system.

The tract base map development will use ArcInfo for all editing work. We will store base maps as ArcInfo coverages and will maintain scanned maps as either TIFF files or ArcInfo grids, depending upon the amount of rectification required to match them to existing tract base files.

The process will proceed as follows:

Clean up of 1990 and 2000 boundary files. We will obtain census tract files in digital form from the Census Bureau. These datasets were derived from the TIGER files by extracting tract boundaries, generating polygon topology, and simplifying lines. We will modify these datasets by further simplifying some lines to eliminate unwanted artifacts such as remnants of piers, removing selected water bodies such as large lakes and coastal bays, and eliminating sliver polygons. Work will proceed on a state-by-state basis.

Creating pre-1990 boundary files. We will project the state datasets and subdivide them into counties. All further work will be done on a county-by-county basis. We will then carry out on-screen editing for pre-1990 census years by county in reverse chronological order starting with 1980. Once a county is complete for all tracted census years, the editor will start a different county. Tract IDs and tract geometry will be edited simultaneously.

Verification. When the county tract base maps are complete for each census year, we will reexamine the datasets for quality assurance. We will employ multiple approaches, including checking tract borders with paper maps, using the Census Bureau tract changes tables to verify changes between census years, and matching tract IDs with identifiers used in Census tables. After the county files have passed quality assurance procedures, we will export the files and archive the original source materials (coverages and scans).

Creating historical attributes. When the tract databases are complete and checked, we will develop the historical attribute information. We will derive information regarding when a tract came into existence, ID changes for the same tract, and “parent” tract and date for split tracts. We will obtain these data through standard geometric overlay functions in ArcInfo.

Constructing sub-county databases. The procedures for minor civil division (MCD) and other sub-county databases are similar to the procedures for census tracts. For each state, we will start with existing datasets and work backward census by census to build boundary compatibility across census years. The sub-county process will diverge from the tract procedures depending upon the type of sub-county unit. We will derive sub-county units with borders coincident with the public land survey system (PLSS) from digital township datasets, and will obtain minor civil divisions by digitizing manuscript maps. Unlike the census tract databases, a significant prerequisite for constructing the MCD-based sub-county databases consists of obtaining appropriate reference materials, and we expect to find that the necessary large-scale paper maps will not be available for all counties in all census years.

Constructing county and state databases. A primary goal is to make the county borders match tract and other sub-county borders for each census year. This means that we will have to use our tract and sub-county datasets when creating county datasets. In most cases, counties will be constructed from a combination of tract, MCD and PLSS data. We will establish an initial database through minor editing of the 1990 and 2000 Census Bureau boundary files, and then work backward to earlier census years. Our primary references for changes in county boundaries will be the Newberry Library’s Atlas of Historical County Boundaries (Denboer 1993-) and a set of Historical United States County Boundary Files created at Louisiana State University (<http://www.ga.lsu.edu/ga/husco.html>). Because of the need to ensure the best fit across geographic entities, we cannot work directly from the LSU files, but they will be a useful supplemental reference where the Newberry maps are not available. Once the county datasets are developed, corresponding state sets will be extracted simply by merging counties.

Interpolation. Changes in boundaries can lead to spatial incompatibility across census years. For example, census tract 101 in 1980 might be split into 101.01 and 101.02 in 1990. The result is that a direct comparison of population statistics, such as median housing value, will not be meaningful unless the data are reaggregated. To address this problem, the NHGIS will allow users to designate a reference year and to estimate population characteristics for all years based on the geography of the reference year. These estimates will be based on simple spatial interpolation incorporated in look-up tables. For example, Enumeration Unit A in 1970 might be spatially equivalent to 0.2 unit A, 0.4 unit B and 0.4 unit C in 2000. These weights will then be applied in comparing population values between the two years.

3. Data Access Tools

We will distribute data and documentation free of charge on the Internet through an integrated data access system. The NHGIS will provide a rich environment for identifying, accessing and using data. Users will extract customized subsets of both data and documentation tailored to their particular research questions. This will not, however, be simply a data extraction system. Rather, it will be a set of tools for navigating documentation, defining datasets, constructing customized variables and adding contextual information. Instead of presenting the data in isolation, this system will provide the user with both the data and a rich informational context within which to interpret it. The electronic boundary files will be seamlessly integrated into the system. Users will have the option of defining their geographic areas of interest through a map interface and will be able to display their results in the form of a table or map. The data access component of the project is substantial, comprising almost twenty percent of our total costs.

The NHGIS data access system represents the next generation of web-based data dissemination. Current tools for accessing aggregate data on the Internet are difficult to navigate and provide only rudimentary topical search features. Such systems usually provide only one method for selecting variables (<http://govinfo.kerr.orst.edu>; <http://www.oseda.missouri.edu/mscdc/apps.html>). They typically require researchers to know technical terminology and provide little or no contextual documentation (<http://ferret.bls.census.gov:80>). Moreover, since existing extraction tools cannot handle the largest and most complex datasets, the assistance of an expert data archivist is still required for many applications. The NHGIS data access tools will take advantage of recent developments in information technology to transcend these limitations.

Variable search and selection. Variable selection poses special challenges for the aggregate data files because the number of variables is extremely large; we estimate the total number of variables in the NHGIS will exceed 300,000. The standard method for variable selection in web-based extraction systems forces users to select variables from a set of static lists. This approach is completely impractical for the NHGIS. Access to aggregate data is further complicated because a different array of subject variables is available at each level of geography in each census year. For example, income and education are available at the tract level in most census years, but not at the block level. Some variables are available in all census years, and others appear only once. Therefore, the development of tools to allow users to locate the subject information they need is of paramount importance.

The NHGIS system will develop a variety of methods for locating variables. Novice users will be offered an expert-systems interview, which will pose a series of questions to identify their areas of substantive interest, the level of geographic detail they need, and the chronological scope of their investigation. More advanced users will be able to search all documentation by variable name, keyword, or variable description, or to drill down from broad subject classifications to specific groups of variables. Users will be able to specify the level of geographic detail and census years at the outset thereby restrict the search universe to variables that meet these requirements. Alternatively, they will be able to explore the tradeoffs between geography and variable availability dynamically. Once they have narrowed their search to a manageable group of variables, the system will highlight the subset of variables that are available for any given combination of geographic level and census year. When they modify the geographical and chronological specifications, the highlighted group of variables will change accordingly. Users will then be able to add variables to a “data basket” for later analysis. At any time during the selection process, users will be able to click on a variable to get a full description and analysis of comparability problems across census years, enumerator instructions, and so on.

The NHGIS will also support several methods of geographic case selection. Most users will specify geography through a clickable map interface. They will first define one or more states, metropolitan areas or regions by clicking on a national map. Then they will specify any finer geographic level of interest—counties, cities, minor civil division, tracts, zip codes, or blocks. This will yield a clickable map with the appropriate boundaries. Users will add geographic selections to a data basket, and at any time will be able to shift to a different geographic level or region. Unlike existing geographic selection methods, users will be able to mix and match different geographies. For example, they will be able to select a combination of municipalities, minor civil divisions, and census tracts within the same extract. Advanced users who know the names of the geographic units of interest will have the option of selecting geographic units from a structured list or using a search engine to locate geographic entities by name or Federal Information Processing System (FIPS) code. The system will allow users to go back at any time and change their geographic and subject variable selections, and will permit users to define their selections in whatever order they wish.

Retrieving and managing data. Most census statistics consist of raw counts of the number of individuals or households in a geographic area with a particular characteristic or combination of characteristics. These data are difficult to interpret in isolation. For example, the raw datasets report the number of Hispanics below the poverty line in each particular census tract, but that statistic is of little use without also knowing the total number of Hispanics in the tract. Accordingly, the NHGIS data access system will provide a variety of statistical tools to allow analyses that are more meaningful than raw population counts. For example, the system will calculate percentages and offer suggestions of the appropriate denominator. In the case of Hispanics below the poverty line, the system would suggest calculating either the percentage of all Hispanics who are below the poverty line or the percentage of all persons below the poverty line who are Hispanic. Similarly, the NHGIS will provide options to calculate means and ratios of variables, such as the ratio of male to female income within census tracts. It will also offer basic arithmetic functions, such as summing across geographic units and across variables (e.g., age groups or ethnic groups). There will be special features to facilitate comparisons across time, such as adjustment of economic data to reflect changes in the Consumer Price Index. Finally, the NHGIS will calculate geographic density measures, such as automobiles per square kilometer.

The system will provide results in the form of tables or thematic maps displayed on the screen. On-screen tables will be self-documented; hyperlinks on each variable and value label will lead to information on sources, universe, comparability, enumerator instructions and other ancillary documentation. We will develop table-formatting software specifically for this project, and will use the ArcIMS tools developed by the Environmental Systems Research Institute (ESRI) to generate maps. Users will also be able to extract datasets for downloading and further local processing. Downloadable statistical datasets will be created as delimited or column-format files, and will be accompanied by customized documentation and data definition files for SAS, SPSS, or Stata. Electronic boundary datasets will be distributed in ArcInfo exchange format ("e00" files), ESRI shape files and MapInfo exchange format ("mid/mif" files).

The procedures for defining a given table, map, or downloadable file can be quite complex, involving variable and geographic selections, statistical function, and table or map formatting. We will allow users to save and retrieve a profile for each analysis they create. This feature will enable users to recreate or modify a table or map made in a previous session.

Using DDI technology for data access. What makes all this possible is the DDI metadata standard. The entire data access system will be driven by the DDI. The DDI underlies both access to documentation and the procedures for locating and extracting statistical information. The NHGIS will include hundreds of data files and hundreds of thousands of variables and geographic entities. A system of this complexity would be impractical if each dataset and variable required customized software development. The data access system will use the same software code for all datasets and variables, relying on the machine-understandable documentation to control both the user interface and data extraction.

Because it is driven by the DDI, our approach eliminates the need to alter the user interface or the underlying extraction engine as the database grows. Adding new data and documentation will be a simple matter of creating new DDI-compliant metadata. It will also be possible to apply the software to aggregate data from other countries with only modest adjustments.

We are near completion of a pilot study for the NHGIS data access system funded by the Minnesota Digital Libraries project. This pilot project, entitled the Public Data Access System (PDAS) automates access to aggregate census data from Summary Tape File 4 (STF4) of the 1990 census. STF4 is one of the most daunting but useful datasets in the Census repertoire. These files provide highly detailed social and economic information for specific racial and ethnic groups at fine geographic levels. For example, they allow researchers and local planners to examine data regarding Hispanics at the tract level within metropolitan areas or immigrant groups such as the Hmong population at the city level. Because of its vast amount of information, the large number of variables it contains, and the corresponding complexity of the dataset, only a few published studies have made use of STF4 (Logan and Alba, 1993). The MRDC found that use of these data expanded dramatically when we made them accessible to researchers using a semi-automated precursor of the PDAS system.

Since PDAS is based on a web-based expert-systems interview, users need not consult a data archivist to make informed use of the data. The interface explores the XML-tagged metadata for appropriate topic and table identification. This activates the current set of programs that extract data and create standardized output. The output of the system is a printable postscript file language that can be transferred back to the web client.

The web server and the extract engine are split into two distinct processes that communicate with the "xml-rpc" protocol. The extract engine processes queries from the user interface, searches DDI compliant XML-tagged codebooks, and then provides feedback to the user interface or the processing system in the format needed. It acts as the intelligent 'black box' interface to the metadata. The PDAS approach allows the use of a DDI codebook as a nodal hub within a network of documents, including electronic boundary files, methodologies, guides, and ancillary documents.

Implications for data access infrastructure in the social sciences. Many social science data providers have already expressed interest in applying our proposed system to their data access needs. The California Digital Library project, California Counts, sees the development of these metadata and access systems as a "tremendous benefit" to their project. The Social Sciences Information Centre–Bonn has already begun discussions on how to use the underlying extract engine as a basis for developing high-level interface systems to heterogeneous collections of data, text and images. Data access system developers in a broad spectrum of other archives and libraries have expressed interest in working with our extraction engine and the DDI to provide aggregate data. These organizations include the Census Bureau, the Centers for Disease Control, Statistics Canada, Networked Social Science Tools and Resources (NESSTAR), NIWI-Steinmetz Institute and the University of California–Berkeley Libraries. These organizations are interested in the proposed system because of its potential to transform the landscape of aggregate data archiving. In addition to providing access to high-demand U.S. census data, these software and metadata tools will open access to useful low-demand datasets that often languish in archives only because of the overhead associated with making their information available.

Because of this national and international interest, we are committed to collaborating with leading data disseminators around the world to ensure that our data access system can meet a broad range of needs. Importantly, we have considerable experience with electronic dissemination techniques. We have been distributing census data through the Internet for a decade, and our IPUMS project has served as a model for many social science dissemination sites. The system we are planning for the NHGIS, however, is entirely new. The challenges posed by the world's largest social science database—with hundreds of thousands of variables in a map-based graphical interface—demand innovative approaches to the problem of electronic dissemination.

4. Project Management and Responsibilities

The complexity of the endeavor is substantial: we must acquire and clean approximately 670 gigabytes of data, create the equivalent of approximately 50,000 pages of DDI-compliant documentation, create 630,00 geographic polygons, write an estimated 60,000 lines of software code, and coordinate activities with our domestic and international collaborators. Accordingly, tightly integrated management is essential.

The principal investigator, chief cartographer and four co-principal investigators will work closely together, with weekly meetings and daily interaction. Although all six investigators will share responsibility for the entire project, each will focus on a different aspect of project management.

- John S. Adams is former Director of the Hubert H. Humphrey Institute of Public Affairs and chair of the Department of Geography at the University of Minnesota. A leading population and urban geographer, Adams has over three decades of experience with census summary data and has directed many large database and cartographic projects. Adams will be in charge of overall project coordination and will be chiefly responsible for the cartographic component of the database.
- Robert B. McMaster has expertise in digital and analytical cartography and urban-based geographic information systems. He will be responsible for the analytical aspects of the cartographic database creation, for developing methods for statistical comparison among enumeration units, and for quality control of the cartographic database.
- Mark Lindberg is Director of the Cartography Laboratory in the Department of Geography. He will be responsible for overseeing the creation of the cartographic base files, for refinement of methodology as the project progresses, and for map database management.
- Wendy Treadwell is Coordinator of the Machine Readable Data Center at the University of Minnesota and President of the Association of Public Data Users. She designed the extensions of the Data Documentation Initiative metadata standard that allow it to accommodate aggregate data. Treadwell will oversee the development of machine-understandable documentation and participate in the development of data access software.

- William Block is Director of Software Development and Research Associate at the Minnesota Population Center. In collaboration with Treadwell, he developed the extraction engine for the Public Data Access System. Block will lead the development team for the NHGIS data access system and will assist Treadwell with the development of metadata.
- Steven Ruggles is Director of the Minnesota Population Center and Distinguished McKnight University Professor at the University of Minnesota. He was principal investigator of the IPUMS project and of separate projects to create national samples of the 1850, 1860, 1870, 1880, 1900, 1910 and 1920 censuses. He will work closely with the programmers building data access tools and with the subcontractors who are preparing the new historical census data.

5. Subcontracts and Collaborators

Although our core staff has broad experience with historical census data, we are turning to the leading national experts in the field to help us compile the most comprehensive possible series of aggregate data for the United States. Andrew Beveridge of Queens College-CUNY will be responsible for digitizing all tract-level data before 1960 that exist solely in paper form. Beveridge is the most experienced scholar now working with historical tract level data and has already converted all historical tract data for New York City into machine-readable form. Michael Haines, the Banfi Vintners Distinguished Professor of Economics at Colgate University, is the foremost authority on historical census data pertaining to states, counties, cities and minor civil divisions. He will fill gaps in those data series by digitizing data from published sources. Myron Gutmann, Director of the Population Research Center at the University of Texas-Austin and Principal Investigator of the Great Plains Project to study population and environment, will contribute machine-readable, county-level data and documentation from the agricultural censuses between 1850 and 2002. Finally, Todd Gardner, a Census Bureau Statistician and expert on historical census geography, will consult on both data and cartographic issues.

We will also collaborate with leading data archives around the world. The Inter-university Consortium for Political and Social Research has agreed to archive, preserve and disseminate both the database itself and the data access tools of the NHGIS. We will collaborate with a broad range of data archives on design of the data access system, to ensure that it is as broadly applicable as possible. These organizations include the Census Bureau, the California Digital Library Project, the Social Sciences Information Centre-Bonn, Statistics Canada, Networked Social Science Tools and Resources (NESSTAR), the NIWI-Steinmetz Institute, and the University of California-Berkeley Libraries.

6. Schedule of Work and Deliverables

The three major project components will proceed on separate but coordinated tracks. During the first year of the project, the Data and Documentation group will acquire, clean, and harmonize the datasets, streamline the codebook mark-up procedures, refine our XML authoring tools and begin marking up datasets. Most of the DDI-compliant metadata will be created during the second and third years of the project, which will allow early access to the datasets and thorough testing. In the final two years, the data and documentation team will focus on the preparation of ancillary documentation, although markup will continue for the later data products of Census 2000 and datasets prepared by our subcontractors.

The mapping team will spend the first year developing county, tract and sub-county boundaries for a small but diverse set of counties from across the country. This test county subset will allow us to refine our procedures and anticipate technical problems before they arise. We will then proceed systematically through the tract, sub-county and county boundary files in that order.

The data access group will spend the first project year generalizing and expanding the capabilities of the PDAS prototype extraction engine. We will then turn to the user interface and develop a functional system for access to statistical data by the end of the second year. During the third year, the data access team will integrate the geographic interface and display features. In the final two years, we will implement the advanced data management and retrieval facilities.

We will release data, documentation, maps and software continuously during the course of the project. We will provide functional access to several of the most important data files—albeit through a rudimentary interface—by the end of the second year of the project. We will then add datasets continuously throughout the remainder of the project as they are completed.

7. Preservation and Sustainability

Long-run survival of the database beyond the project period is critical. The University of Minnesota Population Center guarantees to maintain the system for a period of at least ten years beyond the end of the project. To ensure that the database will be permanently accessible, we have contracted with the Inter-university Consortium for Political and Social Research (ICPSR) to provide long-run preservation and dissemination. The subsidy to ICPSR is essential in view of the size of the task. As Halliman Winsborough, Acting Director of ICPSR, put it when we first broached the issue, "You are proposing as much data as is in our whole archive!" (Personal communication, 6/22/00).

8. Evaluation

We will begin making data available to the public in the second year of the project. From then onward, we will carry out a thorough evaluation of the project's effectiveness at reaching the national academic community and the public. The two basic criteria in this assessment will be the quantity of data distributed and the number of users of the data access system. A third criterion in our self-assessment will be the number of publications that use the database and the number of citations they generate. Because of the length of the publication cycle, there will be a considerable delay before new work generates citations, but by the third year after the data access system goes online we expect to see growing evidence of usage in the citation record.

Although the data access system will be available free of charge, we will require users to complete an on-line registration form to gather information on applications of the database and on user characteristics such as academic discipline and educational status. These data will be used for statistical purposes only and will remain confidential. We will create a log file to record each step researchers take, whether they are browsing documentation or extracting data. We will then analyze these log files statistically to assess patterns of use and to reveal which aspects of the data access system and the documentation deserve the greatest attention.

The development of our data access system is a collaborative enterprise with the user community. We will seek feedback continuously, and plan to carry out periodic surveys of users. After every hundred queries, each user will be asked to fill out a questionnaire regarding their current research applications for the data, the strengths and weaknesses of the access system and the need for new data access capabilities and datasets. We will also obtain reactions at academic and professional conference presentations. When our assessment shows under-utilization by particular disciplines or groups who seem logical clients of the database, we will target the relevant conferences, journals and list servers to advertise the availability and potential of the data.

9. Results of Prior NSF Research

Our most closely related prior NSF research project is the Integrated Public Use Microdata Series (IPUMS). Among several NSF awards for the project, the most important was the first: "Integrated Public Use Microdata Series," SBR-9118299, \$464,913, 4/1992-10/1995. In a sense, the IPUMS project did for census microdata much of what we are proposing to do for census aggregate data. Before the IPUMS, it was difficult to use census microdata in time series because of variations in classification systems, file formats and documentation. The IPUMS transformed a diverse collection of census microdata files into a coherent series of individual-level U.S. census data drawn from thirteen census years between 1850 and 1990. By putting all the census samples in a compatible format with consistent variable codes and integrating their documentation, the IPUMS has greatly simplified the use of multiple census years. Just as important, the IPUMS project pioneered methods of electronic dissemination that have democratized access to these resources.

We plan to integrate the NHGIS and IPUMS data access systems to provide one-stop access to all census data. The NHGIS will incorporate all IPUMS geographies, making it easy for users to add aggregate contextual variables to census microdata. The two projects are highly compatible, but the NHGIS is substantially more complex. Indeed, compared with the NHGIS, the IPUMS project was relatively small in scale. The NHGIS will include 25 times as much data as the IPUMS, and 600 times as many variables.

Many publications have resulted from this work; for examples, see Fitch and Ruggles 2000; Gardner 1995, 1998, 1999, 2000; Gardner, Sobek and Ruggles 1999; Ruggles 1993, 1994a, 1994b, 1995a, 1995b, 1996a, 1996b, 1997a, 1997b, 2000; Ruggles and Brower 2000; Ruggles, Hacker and Sobek 1995; Ruggles and Menard 1995; Ruggles and Sobek 1995, 1998; Ruggles, Gardner, and Sobek 1996; Ruggles, Sobek, and Gardner 1996; Sobek 1996, 1997; Sobek and Ruggles 1999; Block and Star 1995. For additional publications resulting from the IPUMS project, see <http://www.ipums.org/~pipums/research.html>.

References Cited

- Alba, Richard D., Nancy A. Denton, Shu-yin J. Leung and John R. Logan. (1995). "Neighborhood change under conditions of mass immigration: The New York City region, 1970-1990." *International Migration Review* 29: 625-56.
- Arbes, S.J. Jr., A.F. Olshan, D.J. Caplan, V.J. Schoenbach, G.D. Slade and M.J. Symons. (1999). "Factors contributing to the poorer survival of black Americans diagnosed with oral cancer." *Cancer Causes & Control* 10: 513-23.
- Barkley, David L., Mark S. Henry and Shuming Bao. (1998). "The role of local school quality in rural employment and population growth." *Review of Regional Studies* 28: 81-102.
- Becker, K.M., G.E. Glass, W. Brathwaite and J.M. Zenilman. (1998). "Geographic epidemiology of gonorrhea in Baltimore, Maryland, using a geographic information system." *American Journal of Epidemiology* 147: 709-16.
- Block, William C. and Dianne L. Star. (1995). "Data entry and verification." *Historical Methods* 28: 63-5.
- Clark, William A.V. (1998). "Large-scale immigration and political response: Popular reaction in California." *International Journal of Population Geography* 4: 1-10.
- DenBoer, Gordon. (1993-). *Atlas of Historical County Boundaries*. Edited by John H. Long. New York: Charles Scribner's Sons (49 Volumes).
- Denton, Nancy A. and Douglas S. Massey. (1991). "Patterns of neighborhood transition in a multiethnic world: U.S. metropolitan areas, 1970-1980." *Demography* 28: 41-63.
- Duncan, Otis Dudley. (1957). *The Negro population of Chicago; a study of residential succession*. Chicago: University of Chicago Press.
- Farley, Reynolds and William H. Frey. (1996). "Latinos, Asian, and black segregation in U.S. metropolitan areas: Are multiethnic metros different?" *Demography* 33: 35-49.
- Fitch, Catherine A. and Steven Ruggles. (2000). "Historical trends in marriage formation." In Linda Waite, Christine Bachrach, Michelle Hindin, Elizabeth Thomson and Arland Thornton, eds., *Ties that Bind: Perspectives on Marriage and Cohabitation*. Hawthorne: Aldine de Gruyter, 59-88.
- Gardner, Todd. (1995). "Software development for the Public Use Microdata Samples." *Historical Methods* 28: 59-62.
- Gardner, Todd. (1998). "The metropolitan fringe: Suburbanization in the United States before World War II." Ph.D. Dissertation, University of Minnesota.
- Gardner, Todd. (1999). "Metropolitan classification for census years before World War II." *Historical Methods* 32: 139-50.
- Gardner, Todd. (2000). "The slow wave: The changing residential status of cities and suburbs in the United States, 1850-1940." *Journal of Urban History* (forthcoming).
- Gardner, Todd, Matthew Sobek and Steven Ruggles. (1999). "The IPUMS data extraction system." *Historical Methods* 32: 119-24.
- Gutmann, Myron P. (2000). "Scaling and demographic issues in global change research: The Great Plains, 1880-1990." *Climatic Change* 44: 377-39.
- Lang, David M. and Marcia Polansky. (1994). "Patterns of asthma mortality in Philadelphia from 1969 to 1991." *New England Journal of Medicine* 331: 1542-6.
- Latkin C., G.E. Glass and T. Duncan. (1998). "Using geographic information systems to assess spatial patterns of drug use, selection bias and attrition among a sample of injection drug users." *Drug & Alcohol Dependence* 50: 167-75.

- Leclere, Felicia B., Richard G. Rogers and Kimberley Peters. (1998). "Neighborhood social context and racial differences in women's heart disease mortality." *Journal of Health and Social Behavior* 39: 91-108.
- Lieberson, Stanley. (1963). *Ethnic Patterns in American Cities*. Glencoe, IL: The Free Press of Glencoe.
- Liu L., D. Deapen and L. Bernstein. (1998). "Socioeconomic status and cancers of the female breast and reproductive organs: A comparison across racial/ethnic populations in Los Angeles County, California (United States)." *Cancer Causes & Control* 9: 369-80.
- Logan, John R. and Richard D. Alba. (1993). "Locational returns to human capital: Minority access to suburban community resources." *Demography* 30: 243-69.
- Logan, John R., Richard D. Alba, Tom McNulty and Brian Fisher. (1996). "Making a place in the metropolis: locational attainment in cities and suburbs." *Demography* 33: 443-53.
- Massey, Douglas S. and Mitchell L. Eggers. (1990). "The ecology of inequality: minorities and the concentration of poverty, 1970-1980." *American Journal of Sociology* 95: 1153-88.
- Massey, Douglas S. and Nancy A. Denton. (1988). "Suburbanization and segregation in U.S. metropolitan areas." *American Journal of Sociology* 94: 592-626.
- Massey, Douglas S. and Nancy A. Denton. (1998). "The elusive quest for the perfect index of concentration: reply to Egan and Weber." *Social Forces* 76: 1123-34.
- Mielke H. W., C.R. Gonzales, M.K. Smith and P.W. Mielke. (1999). "The urban environment and children's health: soils as an integrator of lead, zinc, and cadmium in New Orleans, Louisiana, U.S.A." *Environmental Research* 81: 117-29.
- Miles-Doan, Rebecca. (1998). "Violence between spouses and intimates: does neighborhood context matter?" *Social Forces* 77: 623-25.
- Miller, David W. and John Modell. (1988). "Teaching United States History with the Great American History Machine." *Historical Methods* 21: 121-34.
- Nuorti, J.P., J.C. Butler, L. Gelling, J. L. Kool, A.L. Reingold and D.J. Vugia. (2000). "Epidemiologic relation between HIV and invasive pneumococcal disease in San Francisco County, California." *Annals of Internal Medicine* 132: 182-90.
- Pulido, Laura. (2000). "Rethinking environmental racism: white privilege and urban development in Southern California." *Annals of the Association of American Geographers* 90: 12-40.
- Ruggles, Steven. (1993). "Historical demography from the census: applications of the American census microdata files." in Roger Schofield and David Reher (eds.) *Old and New Methods in Historical Demography*. Oxford: Oxford University Press, 383-93.
- Ruggles, Steven. (1994a). "The origins of African-American family structure." *American Sociological Review* 59: 136-51.
- Ruggles, Steven. (1994b). "The transformation of American family structure." *American Historical Review* 99: 103-28.
- Ruggles, Steven. (1995a). "Sample designs and sampling errors in the Public Use Microdata Samples." *Historical Methods* 28: 40-6.
- Ruggles, Steven. (1995b). "Family interrelationship coding in the Integrated Public Use Microdata Series." *Historical Methods* 28: 52-8.
- Ruggles, Steven. (1996a). "The effects of demographic change on multigenerational family structure: United States whites 1880-1980." In Alain Bideau, A. Perrenoud, K. A. Lynch and G. Brunet, eds., *Les systèmes démographiques du passé*. Lyons: Centre Jacques Cartier, 21-40.
- Ruggles, Steven. (1996b). "Living arrangements of the elderly in America, 1880-1980." In Tamara K. Hareven, ed., *Aging and Generational Relations Over the Life Course: A Historical and Cross-Cultural Perspective*. New York: Aldine de Gruyter, 254-71.

- Ruggles, Steven. (1997a). "The rise of divorce and separation in the United States, 1880-1980." *Demography* 34: 455-66.
- Ruggles, Steven. (1997b). "The effects of AFDC on American family structure, 1940-1990." *Journal of Family History* 22: 307-25.
- Ruggles, Steven. (2000). "Living arrangements and economic well-being of the aged in the past." *Population Bulletin of the United Nations* (forthcoming).
- Ruggles, Steven, J. David Hacker and Matthew Sobek. (1995). "Order out of chaos: General design of the Integrated Public Use Microdata Series." *Historical Methods* 28: 33-9.
- Ruggles, Steven and Susan Brower. (2000). "New estimates of household composition in the United States, 1850-1990." Forthcoming in Michael Haines, Richard Sutch and Susan Carter, eds., *Historical Statistics of the United States: Millennial Edition*. Cambridge University Press.
- Ruggles, Steven, Todd Gardner and Matthew Sobek. (1996). "Disseminating historical census data on the World Wide Web." *Iassist Quarterly* 20, 4-18.
- Ruggles, Steven and Russell R. Menard. (1995). "The Minnesota Historical Census Projects." *Historical Methods* 28: 6-10.
- Ruggles, Steven and Matthew Sobek. (1998). *IPUMS-98: Integrated Public Use Microdata Series* (five volumes). Minnesota Population Center.
- Ruggles, Steven, Matthew Sobek and Todd Gardner. (1996). "Distributing large historical census samples on the Internet." *History and Computing* 9: 145-59.
- Sayegh A.J., R. Swor, K.H. Chu, R. Jackson, J. Gitlin, R.M. Domeier, E. Basse, D. Smith and W. Fales. (1999). "Does race or socioeconomic status predict adverse outcome after out of hospital cardiac arrest: a multi-center study." *Resuscitation* 40: 141-6.
- Sobek, Matthew. (1996). "Work, status, and income: Men in the American occupational structure since the Nineteenth Century." *Social Science History* 20: 169-207.
- Sobek, Matthew. (1997). *Occupational Structure and the Labor Force in the United States, 1880-1990*. Ph.D. Dissertation, University of Minnesota.
- Sobek, Matthew and Steven Ruggles. (1999). "The IPUMS project: An update." *Historical Methods* 28: 102-10.
- South, Scott J. and Kyle D. Crowder. (1997). "Residential mobility between cities and suburbs: race, suburbanization, and back-to-the-city moves." *Demography* 34: 525-38.
- Sucoff, Clea A. and Dawn M. Upchurch. (1998). "Neighborhood context and the risk of childbearing among metropolitan-area black adolescents." *American Sociological Review* 63: 571-86.
- Taeuber, Karl E. and Alma F. Taeuber. (1965). *Negroes in Cities: residential segregation and neighborhood change*. Chicago: Aldine Publishing.
- Woodruff, T.J., D.A. Axelrad, J. Caldwell, R. Morello-Frosch and A. Rosenbaum. (1999). "Public health implications of 1990 air toxics concentrations across the United States." *Environmental Health Perspectives* 107: A546-8.
- Wyly, Elvin K. and Daniel J. Hammel. (1998). "Modeling the context and contingency of gentrification." *Journal of Urban Affairs* 20: 303-27.